Cultural Transmission and Inductive Biases in Populations of Bayesian Learners

9796003
MSc Speech & Language Processing
The University of Edinburgh

August 2009

Contents

Abstract

Recent research on computational models of language change and cultural evolution in general has focused on the analytical study of languages as dynamic systems, thus avoiding the difficulties of analysing the complex multi-agent interactions underlying numerical simulations of cultural transmission. The same is true for the examination of the effects of inductive biases on language distributions within the Bayesian Iterated Learning Framework. The aim of this work is to test whether the strong results obtained through analytical methods in this framework also extend to finite populations of Bayesian learners, and to investigate what other effects richer population dynamics have on the results. Small world networks are introduced as a tool to model social structures which are shown to play an important role in the outcome of cultural transmission processes. The assumptions behind a Bayesian approach to language learning and its implications will be studied and compared to previous models of language change. While studying the effects of populations on convergence rates in the Bayesian model, the role of more complex population settings for the future of Iterated Learning will also be explored.

1 Introduction

In linguistics, computational simulations have been used extensively to examine language change, language evolution, and to build and test models which increase the understanding of the factors involved in these processes [Briscoe, 2002]. After the realisation of the importance of interaction in the shaping of language and the cumulative adaptive processes underlying change, recent years have seen an increased popularity of models borrowing from mathematical frameworks used in population genetics [Mufwene, 2008] or statistical physics [Niyogi and Berwick, 1995]. While these frameworks are explicitly built to handle population dynamics, the popularity of models that can be studied analytically has led to focus on very simplified models of cultural transmission. This work means to extend work on language transmission and competition within the Bayesian framework, which has been used to examine effects of learning biases on the shape of language, to a population setting. Particular emphasis will be laid on the role of population structure in shaping both language and language change, suggesting that a more systematic study of such effects can help to increase understanding of the underlying factors of such processes and thus result in more accurate models of cultural transmission.

The structure of this dissertation is as follows: this section contains a literature review which puts the work presented in later chapters into context, also pointing out aspects of language change which have been largely disregarded in the modelling literature. Section 2 describes a computational model addressing these shortcomings, and gives a detailed account of the model and simulations. Section 3 presents the results and interpretations of the simulations. Two seperate subsections are dedicated to the plausibility of a Bayesian learning algorithm and the consequences of employing one, particularly in a population setting. Section 4 summarises the results and suggests directions for future work in the field.

The subsequent literature review is structured as follows: firstly, section 1.1 introduces the Iterated Learning framework as a general tool to examine the effects of transmission channel properties on the nature of learned traits. This is followed by a brief overview of traditional approaches to describing and explaining language change, particularly from a historical and sociolinguistic perspective, in section 1.2. Emphasis is laid on the importance of different roles of individuals in causing and driving language change in heterogeneous populations. The discussion leads directly over to section 1.3, where small world networks are presented as a potential tool to investigate effects of population structure on language change, and an extensive review of existing models is dedicated to find a suitable model for usage in an Iterated Learning Model.

Section 1.4 presents a brief summary of attempts to model language acquisition and change computationally, followed by an overview of modelling work on language evolution in section 1.5. This last section also documents the latest results of Bayesian models of language learning. Similarities and differences between the historical and sociolinguistic approach and that of computational models as well as shortcomings of previous modelling work will be summarised in section 1.6.

1.1 Cultural Transmission and Iterated Learning

All languages spoken in the world today are, like all other cultural traits, products of cultural transmission. In contrast to signalling system found in other species (which are innate and rigid), the signal-sign relationships in human languages are not only arbitrary but also learned. This biologically-endowed, uniquely human capacity to not only acquire arbitrary signals, but also to combine them to form compositional meanings of high complexity, is often referred to as the *language acquisition device* [Chomsky, 1957, Chomsky, 1965].

Apart from investigations on the *biological* evolution of language – the exploration of the origin and development of this genetically transmitted language acquisition device – research has been carried out on the structure and development of human languages as they are transferred from one generation to another as part of a *cultural* evolution of language [Christiansen and Kirby, 2003]. It is suggested that the properties of the transfer and learning process itself affect the structure of languages quite significantly, potentially rendering arbitrary explanations on the basis of a *language acquisition device* or *Universal Grammar* unnecessary.

A useful tool to formally investigate the effects of transfer processes on cultural transmission is the Iterated Learning framework [Kirby, 2001, Brighton and Kirby, 2001]: in Iterated Learning Models (ILMs) the (linguistic) output of a population of agents is used as learning input for new agents, which again produce output for the next learning generation. This process can be used in experimental settings with real participants [Kirby et al., 2008, Cornish, 2006] or, as in this work, as the basis for a computational model: the framework itself is very general and can be combined with any kind of acquisition algorithm and arbitrary population and transmission configurations.

The simplest and most well-studied variant of an ILM is that of a *linear transmission chain*, where the population consists of only 1 agent. This agent produces input data for the agent which constitutes the next generation. The simplicity of this model allows computational instantiations of it to be treated analytically by describing the resulting dynamic system as a Markov process. The extensive results of these analyses, in particular of the Bayesian Iterated Learning Model, will be examined in further detail in subsection 1.5.

This simplest version of the ILM is of course only one of a number of imaginable scenarios for investigating cultural transmission. It is largely due to their analytical tractability and experimental feasibility that these linear transmission chain models have been used more extensively in recent studies. This tendency to study only very simplified interaction patterns within populations has been critised both in computational models [Niyogi and Berwick, 2009] as well as in experimental settings [Walker et al., 2009], suggesting that more complex interaction patterns within the framework are necessary to achieve more representative results. Such complex group settings like the replacement method, in which members of a group are continuously replaced by new ones in combination with an ongoing group-based learning task, is itself quite old and has received significant attention in early experimental work on cultural transmission [Mesoudi and Whiten, 2008]. An advantage of these settings is that the usage

of groups not only allows for long-range vertical transmissions (between "generations") but potentially also for horizontal transmission (between members of the same generation) which can lead to population-level agreement dynamics, which are critically missing from linear transmission chain settings. Agreement dynamics have also been of special interest for the question of cultural language evolution, resulting in extensive computational studies in the area, which will also be discussed further in section 1.5.

Since this work aims to enhance the use of complex populations within the Iterated Learning framework, it is worthwhile to consider investigations on the role of population structures in an area of research which has focussed on a particular aspect of cultural transmission, namely the study of language change.

1.2 Language Change

The systematic study of processes of language change originated from philology, and would later be established as its own branch under the umbrella term *historical linguistics*. While early explorations were of a purely descriptive nature and only very speculative on the explanatory level, a set of formal criteria have been established within the discipline to allow for more scientific and insightful investigations of the phenomena at hand [Weinreich et al., 1968].

The spread of language change has often been characterised as a wave-like propagation, where the replacement of an old linguistic variant by a new one follows a logistic growth function, commonly referred to as an S-shaped curve [Kroch, 1989b]. Since it would not be useful for a speaker to acquire a new, rare form over an already established variant, the new variant usually possesses some *functional advantage* over the old one: while the origin of such an advantage varies significantly between the different domains of human language (lexicon, phonetics, syntax, etc), it is often associated with *overgeneralisation* to new contexts. Numerous examples for this exist in various domains of language change, the most well documented and extensively studied examples attesting the transition from Old to Middle to Modern English. These periods exhibited significant syntactic changes, like the rise of periphrastic *do* [Kroch, 1989a] or the shift from an $SOV + V2$ to an SVO syntax in Old English, which is also documented for other languages like Old French, all through the structural *reanalysis* of the input data that learners received [Kroch, 1989b]. Crucially, change does not simply occur due to the *imprecise* learning of speakers [Yang, 2002], which would only result in random fluctuations. Language change is usually grounded in the *reinterpretation* of the input that learners get, and it can be useful to think of language acquisition as a *reconstruction* process. The underlying cause for such reinterpretations is usually a much bigger mystery: if a certain aspect of a language is inherently "unstable" and prone to be replaced by another variant, why hasn't it already happened? The crucial question of *triggering*, also coined the *actuation problem* has been around for decades [Weinreich et al., 1968] and it does not appear like a satisfactory general answer can be given. Indeed, the dependence of such a functional advantage on a massive number of other properties of any specific language under examination renders the feasibility of a general theory of language change unlikely.

But since change originates in learning or other reproduction processes like accommodation [Trudgill, 1982, Kerswill and Williams, 2000], a partial answer to the actuation problem can be found in the nature of the input that learners receive as their *primary linguistic data* (PLD) [Yang, 2002]. While it is clear that no two speakers ever receive the same PLD, the actual question is where shifts in the language input strong enough to trigger a *significant reorganisation* stem from.

While a number of cases of language change could be "explained away" by massive external influences like language contact, it is obvious that even languages with highly secluded speaker communities change over time. Languages are held in constant flux by the omnipresent changes in heterogeneous inputs that each language learner receives, resulting in different inferences about what the language looks like which, in combination with other sources of individual variation, are ultimately manifested in each individual's *idiolect*. The heterogeneous input is on one hand due to the diversity of the sets of speakers that each individual receives their linguistic data from, but on the other hand a significant amount of variation can be found even within single speakers [Chambers and Trudgill, 1980].

Efforts on a more systematic study of other "external", predominatingly social factors involved in language change are relatively recent. Such *ecologies* [Mufwene, 2008], which might subsume a variant's statistical frequency, semantic transparency, regularity, salience and social status of the model speakers as well as more intricate details of social and communicative patterns within a population enable a much more powerful approach to explain language change [Croft, 2000]. All these factors influence the exact distribution of primary linguistic data available during language acquisition [Yang, 2002], which ultimately lead to language change.

Research on such general frameworks is supported by investigations of the qualitative effects of communication patterns within populations on the *rate* of change, they are also suggested to have significant effects on the *structure* of languages spoken within populations [Wray and Grace, 2007]. Recent investigations have reveiled significant correlation between demographic factors of language communities and the subjective structural *complexity* of their respective languages [Lupyan and Dale, 2009]. These results again highlight the importance of the influence of population structure on the outcome of language acquisition.

Since the rate as well as to some degree the direction of language change seem to be determined by demographic properties, it is worth investigating models of social networks to study such apparent influences effectively. We will thus have to leave it with this indeed only very rough outline of the concepts and issues in the study of language change to focus on the social aspect of language transmission, but will return to an outline of the use computational models to study language change in section 1.4.

1.3 The Strength of Weak Ties and Small World Networks

The increasing interest in the subtle *external* factors involved in potentially triggering language change has emphasised the role of social aspects of communication and language trans-

mission. A major factor in characterising the communicative patterns within a population lies in the nature of the social networks in which speakers are embedded and interact.

General observations about the origin and spread of linguistic innovations (at least in modern societies) are in fact not new: early sociolinguistic enquiries have suggested that the introduction of new variants into a language often stems from individuals which are found at the periphery of the social network of a speaker community, rather than from core members [Milroy and Milroy, 1985]. Since these less central members tend to have more contacts outside the community it is not surprising that they would be more likely to introduce change. On the other hand such members are often not *influential* enough in the sense that a change introduced by them would be readily adapted by the rest of the speaker community. The fact that a variant will only spread given it is adopted by more influential core members, subsequently making *infection* of the rest of the community significantly easier, sketches a rough outline of the stages through which any change usually proceeds and highlights the very different roles of individuals in the process [Milroy and Milroy, 1985].

Observations like these carry a strong resemblence to examinations of the spread of diseases, innovations or information flow in general, which emphasise the distinct roles of individuals who are found at the tighly knit core of a community, or less centrally situated ones which conform less to the prototypical properties of a certain group. The importance of these less central members as potentially *bridging* to other tightly knit group clusters had been long suspected [Milgram, 1967] before it was first studied empiricially. While most social interactions take place in small, dense and seemingly separate communities, the impression of a *small world* where everybody seems to know everybody else over just a few intermediate steps is omnipresent in modern society. This intuition was confirmed and quantified in the famous *Small World Experiment* [Travers and Milgram, 1969], where subjects across the United States were instructed to send a letter to a target person in a different city who they didn't know, but only by forwarding it to another person they knew on a first name basis who they thought would be closer to the target person. The average of 5 steps required to reach the target highlighted the importance of these long-distance connections and the humans' ability to effectively make use of them, the effects of which were later also studied qualitatively and developed into theoretical frameworks [Granovetter, 1973, Granovetter, 1983].

Since the different roles of individuals in the spread of linguistic and other innovations seems to be an important aspect of language change, qualitative studies in this area are therefore not only relevant to increase the understanding of the factors involved in processes of language change as such: much like studies of disease spread help in the field of epidemiology [Watts and Strogatz, 1998], insights on general patterns of change could also be of potential interest for areas such as language politics or language planning.

Keeping these methodological motivations in mind, we will now turn to more technical approaches to the study of social networks, particularly computational network models, and see what they have to offer for a more systematic study of the social *ecologies* of language change.

1.3.1 A History of Small World Networks

The advent of the internet, accompanied by easier availability of processing power, enabled the collection and analysis of huge corpora in the area of complex networks and self-organising systems. The degree distribution of websites on the internet itself turned out to be scale free, i.e. to follow a power law: when plotting the in-degree (the number of links pointing to a website) against the number of websites with this in-degree, it would appear as a straight, decreasing line on a log-log plot. These kinds of distributions, which can be accurately characterised by the negative slope γ of the straight line, were found in other self-organising systems like biological food networks or scientific authorship. The networks' strict adherence to global statistical properties without a central designer overseeing the system's development demanded new, formal theories for the development of such networks.

An avalanche of publications on ever more natural networks exhibiting the same properties soon revealed that the strictly scale-free networks were in fact only a subset of a type of networks that could be described by a very intuitive property: strong local clustering of nodes in combination with a low average shortest path length between all pairs of nodes, achieved by numerous random long-distance shortcuts between dense clusters. The degree of clustering can be described by the clustering coefficient C, which measures the degree of *transitivity* within a network: the number of triangles in the network graph divided by the number of connected triples of vertices gives a measure of how likely it is that two nodes are connected to each other, given that they share a common neighbour [Newman and Park, 2003]. This clustering coefficient, which was found to be significantly higher than what would be expected from random networks, where it approximately follows $C \approx \frac{\langle k \rangle}{n}$ [Watts and Strogatz, 1998], together with a low average shortest path length, growing proportional to the logarithm of the total network size n, establishes these networks as some kind of "golden middle" between the extremes of regular networks (high clustering but long average pairwise distance) on one and random networks (low average pairwise distance but very low clustering) on the other hand [Watts and Strogatz, 1998]. After Milgram's "Small World" experiment this type of network was coined *Small World Networks*.

A number of models have been proposed to account for the uniform global structure of small world networks. Some of the algorithms, which try to give an intuitive understanding about the kinds of local decisions that are required to yield such complex global behaviour, will be presented here. Since small world networks are a potential tool to investigate the effects of different roles of individuals in a heterogeneous population, particular attention will be paid to features of the algorithms which would make them suitable for extensive analysis of the effects of population structures on cultural transmission.

The earliest model by [Watts and Strogatz, 1998] was based on the *rewiring* of a completely regular network: starting from a ring network, random shortcuts would be introduced to lower the average pairwise distance. While the model could successfully interpolate the continuum between completely regular on one and completely random networks on the other hand, it only exhibited an exponential degree distribution (which differs from a scale-free

distribution in that it does not have a "fat tail", meaning that the degrees of a few nodes can get arbitrarily large) and also failed to give an explanation for the source of that particular structure: real networks exhibit continuous growth and are not modified or "rewired" versions of regular networks.

This shortcoming was addressed by a number of *growing* network models. Most prominently the *BA model* by [Barabási and Albert, 1999] exhibited continuous growth where the *preferential attachment* property governed the addition of new nodes and was thus responsible for the emergent scale-free degree distribution: the connections that a newly inserted node made with pre-existing nodes was influenced by the connectivity of the nodes already in the network, with the establishment of a connection more likely if the other node already had a big number of neighbours. This property, also referred to as the "rich get richer" syndrome, turned out to be a defining feature of technical networks like the distribution of weblinks on the internet, and indeed also the cause for the scale free degree distribution in those kinds of technical networks. It did however fail to account for social and economical networks like the patterns found in movie actor collaboration network: in these kinds of networks there is indeed a cutoff point in terms of the number of neighbours a node can have [Amaral et al., 2000]. In pure preferential attachment models the age of a node and its in-degree are positively correlated, something not observed in many real-world networks.

A solution to this discrepancy was the introduction of a parameter to control a node's *aging*, or otherwise a *maximum degree* that a node could have. The model by [Amaral et al., 2000] adds both these features to the original *BA model*, thus enabling accurate modelling of scale-free distributions with a cutoff for high degrees, producing so-called *broad-scale* distributions.

An alternative approach to achieve the same feat, albeit with a more realistic network development process was suggested by [Bornholdt and Ebel, 2001]: in their model, which is also based on the continuous growth model by [Barabási and Albert, 1999], the artefact of correlated age and in-degree of a node is circumvented by decoupling the insertion of new nodes and new edges. A single relative growth parameter α is used to specify the density of the network, and can successfully reproduce scale free distributions for a range of values for the slope parameter γ.

The ongoing study of real world networks revealed more and more defining properties of different types of real world networks, particularly that social networks are a lot more different from other small world networks than previously thought: while clustering is significantly higher than in random networks for all types of small world networks, the clustering coefficient turned out to be several times, sometimes even magnitudes higher for social networks [Jin et al., 2001].

Realising the inadequacy of the processes responsible for the development of technical networks, a number of models for social network evolution were suggested, focussing on the continuous change of connection patterns within a network of fixed size. Incorporating a *cost* for maintaining a connection with another individual and allowing for connections to be broken through decay processes, these *acquaintance network* models would often replace the preferential attachment property with the more local process of *triadic closure* to achieve

extremely strong clustering: triadic closure is the process through which a triple of connected nodes are brought together to form a triangle. This process is modelled after the idea of one person introducing two of its friends who have not previously met to one another, leading to more transitive linking and thus increasing the level of local clustering [Davidsen et al., 2002].

Other models have been proposed which attempt to make the intuitive idea behind the triadic closure process explicit: that the increase in clustering does not take place at random positions in the network, but that it is guided by an underlying *community structure*, in real world networks often based on properties like geographical divisions, which is responsible for the extreme degrees of clustering [Girvan and Newman, 2002, Jin et al., 2001, Newman and Park, 2003].

Yet another more easily comprehensible property that makes social networks stand out from other types of small world networks is that of assortative mixing: a network is said to be assortatively mixed if the degrees of connected nodes are positively correlated. This turns out to be the case for many kinds of social and economical networks, whereas correlation is usually negative for technical networks [Newman, 2002, Newman and Park, 2003]. An increasing number of models dedicated to the study of social network structures incorporate different measures to achieve assortativity [Toivonen et al., 2006, Newman et al., 2002], although the actual process underlying this feature is not fully understood.

A severe problem for most studies so far lies in the limitation of data to synchronic *snapshots* of the networks: all properties observed in these static datasets are the results of continuous and interactive processes, in-depth examinations of which are still sparse. First analyses of ongoing structural developments in huge corpora of digital communicating systems usage have highlighted the very different behaviours underlying small groups and larger institutions [Palla et al., 2007]. More refined observations of processes like community birth, growth and death in more diverse corpora which are currently being undertaken are likely to result in even more realistic models of social network evolution. Research in the field of small world and particularly social networks is thus ongoing, with the focus further shifting away from static properties to understanding the more interactive aspects of community development [Watts, 2004, Kossinets and Watts, 2006, Watts, 2007].

1.3.2 A Small World Network Model for Iterated Learning

The features of the network models discussed in the previous section are manifold, and an overview of the algorithms and their properties can be found in table 1: while all of them might share the same goal, which is to produce realistic models of small world networks, many of them are in fact designed with different applications in mind, and their properties are therefore designed to meet particular requirements. To answer the question of which of the algorithms would be suitable to model population structures on which an Iterated Learning Model could be based, it is thus necessary to consider which properties such a model should ideally have.

Since the nature of ILMs is based on the continuous addition of new and replacement of

old agents, no matter what complexity of the actual population is, the small world network model has to incorporate a mechanism to add new agents on one hand, but it should also keep its small world property when it is truncated, i.e. when old nodes are removed. For simplicity we will want the input data to be sampled right at the insertion of the agent, so that it can "learn" its language immediately. The addition of new edges should thus take place at the same time as the insertion of the node – the later addition of edges across the network will not have any direct influence on the learning process. This property is particular to the task of language learning, since it is subject to a critical period – for future models incorporating the effects of accomodation or more general cases of cultural transmission this constraint might not be necessary.

Since cultural transmission takes place in social networks, the network model should be oriented towards the features of these networks: apart from the limited age of a node, a maximum in-degree might be a useful constraint as well, although it is unlikely to play a role for smaller network sizes and networks without preferential attachment. Preferential attachment itself has been shown to only occur in technical networks and should thus be avoided in social network models, making models with a Poissonian degree distribution the most likely candidates for a model of Iterated Learning. The final property under inspection is assortativity, which has been shown to be the determining difference between social and more technical networks [Newman, 2002]. While these results suggest that a social network model for iterated learning might want to incorporate assortative mixing, it should be noted that some of the network properties might deviate significantly, even in different types of social networks: most real world analyses are based on readily available digital corpora such as coauthorship or film actor collaborations [Newman, 2002], which do not necessarily reflect the properties of communicative patterns within social networks, particularly regarding infants' sources of primary linguistic data. Moreover, these communicative patterns are bound to vary extensively depending on demographics, and even more so in different cultures. While the more intricate properties of network models and their effect on iterated learning might be an interesting starting point for future research, a preliminary model such as presented in this work cannot incorporate or even accurately determine all the properties which might be desirable for such a model. As can be seen from table 1 no algorithm meets all requirements perfectly, suggesting that compromises will have to be made for the use in this work. Quantitative studies of these actual social networks that will shed more light on which of the features of small world networks are crucial, and might thus yield a more paradigmatic set of conditions that an ideal model for Iterated Learning should fulfill.

Model	Growth	Trunc.	Degree distribution	Parameters	Pref. attachm.	Assortativity
[Watts and Strogatz, 1998]	✗	n.a.	no fat tail	p (Rewiring)	✗	?
[Barabási and Albert, 1999]	✓	?	scale-free ($\gamma = 2.9 \pm 0.1$)		✓	✗
[Amaral et al., 2000]	✓	✓	single-scale (exp/Gaussian)	Aging, k_{max}	✓	✗
[Bornholdt and Ebel, 2001]	✓(incr.)	?	scale-free ($\gamma = 1 + \frac{1}{1-\alpha}$)	rel. growth α	✓	✗
[Jin et al., 2001]	✗	✓(incr.)	sharply peaked	$r_0, r_1, \gamma, k_{max}$	✓	✓
[Newman et al., 2002]	✗	✗	arbitrary		✗	✗
[Newman, 2002]	✗	?	arbitrary		✗	✓
[Davidsen et al., 2002]	✓(incr.)	✓(incr.)	$p \to 0$ scale-free, $p \to 1$ Poisson	p	✗	?
[Toivonen et al., 2006]	✓	✗	scale-free ($\gamma = 3 + \frac{2}{m_s}$)	N_0, m_r, m_s	✗	✓

Table 1: Comparison of small world (and particularly social) network models. The properties under examination are those relevant for the study of cultural transmission: continuous addition of new nodes (growth), continuous removal of old nodes (truncatability), a Poissonian or otherwise peaked degree distribution and lack of preferential attachment. (incr.) signifies that edges are continously added (or removed) all over the network and not only upon insertion of a new (or deletion of an old) node, a property that should ideally be avoided for an Iterated Learning Model of language transmission.

1.4 Computational Simulations of Language Change

Given these insights into a useful tool to study social networks quantitatively, we will now return to the question of language change and see how computational simulations have attempted to tackle open questions in the field, particularly those involving small world networks.

Given the amount of early work on computational models of language acquisition, computational explorations of language change appeared quite late (see [Niyogi, 2002] for a review). The earliest models of processes of language change were still focussed on very formal characterisations of language acquisition, often built around the notion of identification of a *target grammar* "in the limit" [Gold, 1967]. While such studies provided formal proof of the long-established intuition that language acquisition was impossible without prior *innate* knowledge of the structure of possible languages [Niyogi, 2006], the predictive results of such models were far less satisfactory.

Early models of language change which were based on acquisition through computational frameworks such as genetic algorithms [Clark and Roberts, 1993] yielded very strong and rigid predictions about the learnability or stability of certain grammars, often contradicting observations found on languages in the real world. These formal investigations also led to a "logical problem of language change" [Niyogi and Berwick, 1995]: assuming perfect learnability of languages, a necessary assumption given the remarkable language acquisition capabilities of humans, the formal frameworks seemed to be irreconcilable with processes of language change.

The increased usage of psychocomputational models of language acquisition promised more realistic models of change. A prime example for such an acquisition algorithm is the *Triggering Learning Algorithm* (TLA): its solution to an effective acquisition process was the idea that learners attend to certain *triggers* which would help in establishing the correct grammatical "settings" for a language, e.g. within the Principles and Parameters frameworks [Gibson and Wexler, 1994]. The learning process is performed on-line and thus constitutes a significant improvement over previous batch-based algorithms: given an input sentence, the learner would try to parse the sentence with its current grammar. If the grammar fails to account for the structure of the input sentence, not yielding a successful parse, the learner would be "triggered" to randomly flip one (or more) of its current parameter settings. The nature of the algorithm was still oriented towards notions of formal learnability and identification of a *target grammar* "in the limit", enabling extensive analytical investigations of its behaviour in various realistic parameter spaces with different target grammars [Niyogi, 2006]. The TLA was subsequently studied in a number of computational models [Clark, 1997, Kirby and Hurford, 1997] which, however, still suffered from extreme predictions which did not reflect distributions of languages in the real world.

The fact that the outcome of any particular TLA run is obviously highly dependent on the very last training sentences that it receives does not make it very robust, thus also questioning its psychological plausibility. Addressing this precise issue new psychocomputational models

of language acquisition have been proposed, a popular example being the Variational Learning Model [Yang, 2002]. The model attempts to bridge the gap between the strict parametric view on language in frameworks like P&P or Optimality Theory on one and statistical learning on the other hand [Yang, 2004]: it does so by regarding the language learning process as constant competition between all possible grammars, with every input sentence that can be parsed by a certain grammar increasing the degree of belief in all parameter values associated with that grammar. These probability updates are again initiated by triggers, where in this model only unambiguous evidence or *signatures* are treated as triggers [Yang, 2002]. Insights into the way in which *input filtering* seems to happen in human syntactic acquisition [Lightfoot, 1991, Pearl and Weinberg, 2007] are considered as well in order to model the learning process as closely as possible. Crucially, the model's ability to use multiple grammars at the same time enables more realistic modelling of language variation, and it was shown to successfully reproduce patterns of language change observed throughout history based on the reported input distributions of the relevant periods. The Variational Learning Model has also been applied to other domains of language like learning the assignment of words to morphological classes and is successful at predicting a variety of crosslinguistic acquisition patterns.

One of the more extensive early works which directly addresses theoretical issues in the study of language change is [Nettle, 1999b], which attempts to relate the results of simulations on the success of language change in different social settings back to the rates of language change in real human populations [Nettle, 1999a]. The basic result was that a functional bias alone was not sufficient to guarantee the spread and adoption of a new variant on a 2-dimensional grid of agents: language change was only guaranteed when agents assigned different weights to the learning input they got from different neighbours, resembling some sort of "social impact", and a few members of the community had to be assigned extraordinary influence. These results were also related to group size, suggesting that the rate of linguistic change would be faster in smaller and thus more coherent speaker communities, including a higher susceptibility to borrowing [Nettle, 1999a].

[Ke et al., 2008] systematically investigated the effects of different network structures, among them scale-free and small world networks, on the time required for a language change to spread through an entire population. Their results showed that the clustered nature of small world networks and particularly scale-free networks did not only reduce the time required to spread through the population, but that the network also facilitated logistic growth, whereas other network types showed linear growth functions. Upon reduction of the functional advantage of the new variant however, small world networks lost this behaviour and receded to linear, albeit more steady growth, with only random and scale-free networks retaining S-shaped curves, paired with a decreasing probability of a change to actually succeed.

A recent model by [Troutman et al., 2008] further highlights the influence of social networks on the spread of language change. Crucially, their model failed to reproduce the characteristical S-shaped curves unless agents had the possibility to proportionally use both competing variants or grammars rather than strictly sticking to either. Their results also point out how social structure can be a crucial factor in bringing the spread of a linguistic

innovation to a halt, thus causing the formation of dialect subgroups.

At the end of this brief overview of previous attempts to simulate language change it is worth noting that all of the models presented here can be assigned to two distinct camps or approaches to the role of populations in language change [Briscoe, 2000] which are also reflected in approaches in historical linguistics [Bynon, 1977], resulting in very different model predictions. As was already pointed out the formal and often probabilistic models, like most instantiations of the TLA, attempt to tackle the question of language change as an abstract process of a dynamic system of language changing according to (stochastic) rules. These approaches tend to disregard the languages' setting in a population and hence do not run into the same problems as theoretical accounts of language change, such as the actuation problem. As a result they produce very strong predictions which are often either irreconcilable or simply hard to relate to real language. Numerical simulations on the other hand, which usually aim to reproduce particular instances of language change observed in the real world, are not only much more easily interpretable, but consequently also a useful tool to study concrete open questions in this area. While these models are not intended as powerful general theories of language change, they do suffer from the same issues as theoretical treatments of language change and can thus also be used more readily to test potential answers to these questions on a model.

Another noteworthy issue is the origin of S-shaped curves: the various models presented here report very different requirements to reproduce this peculiar property of language change. Particularly the result of [Ke et al., 2008], which documents logistic growth in certain network types whereas it is absent in others using exactly the same learning algorithm suggests that this feature might actually be a property of the population structure in which the change is embedded rather than of the process of change itself.

1.5 Computational Simulations of Language Evolution

Since the late 1980s, there has been a significant amount of computational work on the question of *cultural* language evolution: where (human) language comes from and, particularly, which factors other than the nature of the biological *acquisition device* are responsible for its peculiar and often arbitrary structural constraints. Simulations in this area differ from the more readily applicable simulations of language change in that they usually do not have any real data to reproduce or compare to and are thus of a more experimental and exploratory nature. This, however, does not imply that they are not readily interpretable: studies of Iterated Learning in multi-agent simulations have solved puzzles like the emergence of recursive compositionality simply as a consequence of a bottleneck in cultural transmission [Kirby, 2002] or the origin of word regularity and irregularity [Kirby, 2001].

A number of mathematical models which borrow heavily from frameworks used in evolutionary theory and statistical physics have been proposed. One of them is the family of NB models, which extend the Cavalli-Sforza and Feldman model of cultural transmission, itself inspired by models in population genetics, to language [Niyogi and Berwick, 1995,

Niyogi and Berwick, 1997]. These models of language evolution can not be clearly separated from models of language change and often implicitly regard the two as one and the same thing. Crucially, all learners in NB models receive input data that is drawn from the same distribution determined by the composition of the parent population, ignoring the heterogeneity of linguistic input found in reality. Other analytical models from statistical physics [Schwämmle, 2005] have been adapted to study the dynamics of language systems from the point of different research questions, such as the role of selective reproduction on the stability of linguistic coherence within a population [Komarova and Nowak, 2003]. Again, these models make use of very simplified assumptions about the nature of language and often include analyses of dynamic systems assuming an infinite population size, and are thus hard to relate to observable languages.

Driven by research in the field of *Artificial Life*, simulations on *agreement dynamics* in populations of agents have received particular attention. One very well-studied model of language development is the so-called Naming Game [Baronchelli et al., 2006]. In this *language game* an arbitrarily arranged population of agents has the task to arrive at a globally shared "name" for an object while only being able to communicate to a restricted number of neighbours. At the beginning of the game agents invent random names which are then propagated through the population. Since the transmission is completely noise-free, this results in global agreement on one of these names at some point [Baronchelli et al., 2008]. The dynamics of the model have been studied extensively in the mean field case as well as under a variety of population structures [Dall'Asta et al., 2006b], reveiling a decrease in convergence time as well as a reduced memory load of agents in small world networks [Dall'Asta et al., 2006a].

In a similar fashion the *Utterance Selection Model* by Baxter et al. studies the propagation of innovation through complex networks as an adaptive process: the model means to address the social dimension of language change by incorporating the weighted selection of variants or competing *linguemes* within a population, the ultimate outcome of any simulation run being only a single, most adaptive variant remaining [Blythe and Croft, 2009]. The approach, which focuses on the determination of fixation times of variants, characterises language as a complex adaptive system and applies analytical tools used in statistical physics [Baxter et al., 2006].

A rather new approach in the area of simulations of cultural transmission in an Iterated Learning Model is Bayesian Learning [Griffiths and Kalish, 2005]. The advantage of the Bayesian framework, which has been used extensively to produce accurate models of human cognitive functions and reasoning processes, is that it makes it possible (and necessary) to make the learner's expectations at the learning task explicit. In a classical ILM fashion a Bayesian learner has to select one of a number of competing hypotheses based on the input data it gets from the previous generation. The probabilistic nature of the Bayesian model makes it ideal for analytical investigation by describing the system dynamics as a Markov process: assuming a linear transmission chain with a single agent per generation, the transition probabilities between different system states can be expressed and the overall probability of the system states, i.e. the distribution of languages upon convergence, can be calculated.

In this Bayesian Iterated Learning Model (BILM), which will also be the model adopted later in this work (see section 2.1), an agent choses one out of a number of possible hypotheses (representing languages or grammars) based on the input which it receives from its environment as well as a prior probability or expectation for each of the hypotheses, which is assumed to be innately given. The system behaviour has been studied extensively for the simplest case of two competing grammars, with two different learning strategies, *Sampling from the Posterior* and *Maximum a Posteriori* (MAP): when Sampling from the Posterior, the agent probabilistically selects one of the possible hypotheses according to their respective posterior probability, thus constituting an indeterministic learning strategy. Analysis of this strategy reveils that over time the distribution of hypotheses converges to a *stationary distribution* which is the same as the prior distribution of hypotheses [Griffiths and Kalish, 2005]: once the transmission chain has converged, the probability of an agent in the chain having a certain hypothesis is equal to the prior probability of that hypothesis, suggesting that innate biases would be directly reflected in the observable distribution of cultural traits, particularly language.

The other strategy, MAP, in which an agent deterministically chooses the hypothesis with the higher posterior probability, has been studied less extensively. The formal analysis of this strategy which guarantees a deterministic decision by the agent reveils a stationary distribution which also reflects the prior probability distribution, only that this time differences in the priors are exaggerated, with a slightly favoured hypothesis becoming overproportionally represented in a transmission chain [Kirby et al., 2007, Griffiths and Kalish, 2007].

The formal nature of the BILM is in some respects very similar to earlier modelling attempts on language change, which would also be investigated analytically. The different outcomes of the analytical and population-based approaches have already been pointed out, and the model's behaviour was also studied beyond the simple transmission chain layout. [Smith, 2009] investigated the model's behaviour when relieving the single-data-source constraint: given an infinite number of transmission chains running in parallel and assuming that an agent's decision would be based on the data sampled from 2 members of the previous population, the model exhibited sensitivity to the *initial distribution* of languages in the population. Depending on whether the initial proportion of agents with some hypothesis is on either side of a bifurcation point which depends on the prior distribution as well as the number of sampled datapoints, the system would converge to one of two stable stationary distributions at opposite extremes, i.e. it showed a tendency to adhere to either hypothesis. This unexpected behaviour already hints at the significantly different dynamics of the Bayesian Iterated Learning Model in larger and more complex populations, which is what will be addressed in the rest of this work.

1.6 Summary

The discussion of Iterated Learning Models provided here has highlighted the effects of transmission channels on the outcome of learning processes, and the step away from linear chains to

examining the Bayesian model's behaviour in a complex network is the main issue addressed in the remainder of this work. While questions of global agreement have been the subject of studies before, the BILMs role as a *dynamic* system without a predetermined outcome or fixation is expected to yield interesting results.

Crucially, the stationary distribution of languages predicted by previous treatments of the BILM does not signify a homogenous spread of speakers across a population. Rather than just human learning behaviour, it is the geographical distribution of languages together with social and political influences which is responsible for the propagation or extinction of such traits, a factor which has been left out of studies of the BILM so far.

The two different approaches to the modelling of language *change* have already been noted earlier. The weakness of most mathematical models suggested so far was their tendency to focus on the dynamics of *language* as an adaptive system, thus largely ignoring the populations, which are actually responsible for the triggering of changes [Eckardt, 2008]. The shortcoming of most numerical attempts on the other hand was the mostly deterministic outcome of simulations [Wang and Minett, 2005] – for every specific historical example under investigation a different scenario was constructed to predict the actual outcome. While this approach can be helpful in exploring the structure or causes of specific language change processes, it does not contribute to a generalised theory of language change. The introduction of social networks and investigation of the resulting population behaviour in a truly dynamic and not pre-determined model would not only increase the understanding of the phenomena at hand, but also enable the direct examination of effects like population structure on past and present instances of language change or cultural transmission in general. The model presented in the remainder of this work is intended to bridge the gap between these two approaches. By putting the Bayesian Iterated Learning Model into a population setting it will be possible to study the model's exact predictions about language competition in such complex finite populations.

2 Model Design

This section presents the models and simulations which constitute the core contribution of this work. Section 2.1 describes the Bayesian learning algorithm adopted in this implementation, including a brief discussion of the implications of using it in a population setting. Section 2.2 describes the algorithm which is used to grow and evolve the small world networks in which the agents interact. Finally, the parameter settings inspected in the simulation runs are discussed in section 2.3.

2.1 Learning Algorithm

The learning algorithm adopted for the simulations is based on Bayesian inference and has already been studied analytically for linear transmission chains [Griffiths and Kalish, 2005, Griffiths and Kalish, 2007]. The setup adopted in this work consists of two competing hy-

potheses about some aspect of a language or grammar (referred to as h_0 and h_1 throughout the rest of this treatment) that the learners can adopt with their corresponding prior probabilities which are used for Bayesian inference. The hypotheses themselves might correspond to different settings of a parameter within the Principles and Parameters paradigm, the prior biases representing preferential properties due to processing constraints or similar. Each hypothesis has a corresponding prototypical utterance or data item associated with it. When data is sampled from an agent it is most likely to emit the item corresponding to its hypothesis – the exact emission probabilities are governed by the error term ϵ: to accommodate occasional "slips" and introduce the possibility of change and mislearning in the first place, there is a fixed probability of emitting a "wrong" signal, namely the one associated with the complimentary hypothesis. For the remainder of this work the error term is assumed to be $\epsilon = .05$, as was used in most previous treatments of the algorithm.

The final parameter is actually not part of the language and learning model itself, but rather of the Iterated Learning Model incorporating the algorithm: the bottleneck size b controls how many data samples a learner receives from its parent population in order to infer a hypothesis. Given b data points Bayes rule is used to calculate the posterior probabilities of each hypothesis [Griffiths and Kalish, 2005]:

$$p(h|d) = \frac{p(d|h)p(h)}{\sum_{h'} p(d|h')p(h')}$$

Given the posterior probabilities for the two hypotheses, the learner chooses a hypothesis on the basis of one of two possible inference strategies, both of which will be investigated in the simulations: the strategy of *Sampling from the Posterior* determines which hypothesis to adopt by randomly selecting one based on their posterior probabilities. *Maximum a Posteriori* (MAP) on the other hand chooses whichever hypothesis has the higher posterior probability, and only backs off to random selection given equal posteriors.

2.1.1 Bayesian Rationality in Heterogeneous Populations

The promise of a "rational" decision on the side of the learner (as in making the optimal decision based on the data available to him and his prior beliefs) relies on the crucial assumption that the data is indeed sampled from the hypotheses that the learner has knowledge of. Strictly speaking the two hypotheses considered in the model as adopted here do not fulfill these requirements, since the data will in fact be sampled from multiple agents while the learner only has single source hypothesis at his disposal. This issue has already been addressed elsewhere [Ferdinand and Zuidema, 2009, Smith, 2009] and will be discussed in more detail in section 3.6.

2.2 Network Design

The growth model for small world networks used throughout the experiments is based on the algorithm outlined in [Davidsen et al., 2002]. The main difference to the original model,

which was developed to simulate acquaintance networks and is built on the process of triadic closures, lies in the timing of the addition of new links: whereas with the original algorithm edges were continuously added all across the network, this process is now an integral part of the addition or replacement of a new node. The parameters of the model are as follows:

1. n, the maximum number of nodes in the network
2. m, the maximum number of initial connections a newly inserted node has

Given these two parameters and an initial network consisting of three fully connected nodes, in each iteration the network is updated as follows:

1. if the network already contains n nodes, remove the oldest node
2. Add a new node:
   (a) randomly select 2 *distinct* initial neighbours from the entire network
   (b) perform $m - 2$ triadic closures:
      i. randomly select one of the newly inserted node's current neighbours
      ii. randomly select one of the neighbour node's neighbours (excluding the newly inserted node itself). Establish a connection between this node and the newly inserted node if they are not already connected.

The nature of the algorithm allows for slight variation in the degree of single nodes, guaranteeing a minimum number of two and a maximum of m neighbours *upon insertion*. A node's degree varies throughout its lifetime due to the removal of neighbours as well as creation of new connections through newly inserted nodes.

The development of the network's mean degree $\langle k \rangle$ as well as clustering coefficient C, which lies significantly above the coefficient found for a random network with the same mean degree, can be seen in Figure 1 for two different population sizes.

Notably, both measures are even higher during the initialisation phase of the network, i.e. before the network has grown to its full size n, which takes $n - 3$ iterations. In order to guarantee an even distribution of network properties in the simulations, any newly grown network would undergo another 1000 iterations before being used for modelling purposes.

2.3 Simulations

In the simulations, the Bayesian learning algorithm was run in populations of fully connected networks (i.e. a learning agent might potentially sample input data from any agent in the network) as well as the small world networks created by the algorithm described in the previous section. The goals were on one hand to study the dynamics of the populations and classify patterns of change, and on the other hand to determine potential differences between the behaviours in fully connected and small world networks.

The simulations were carried out using a variety of settings for each of the following parameters:

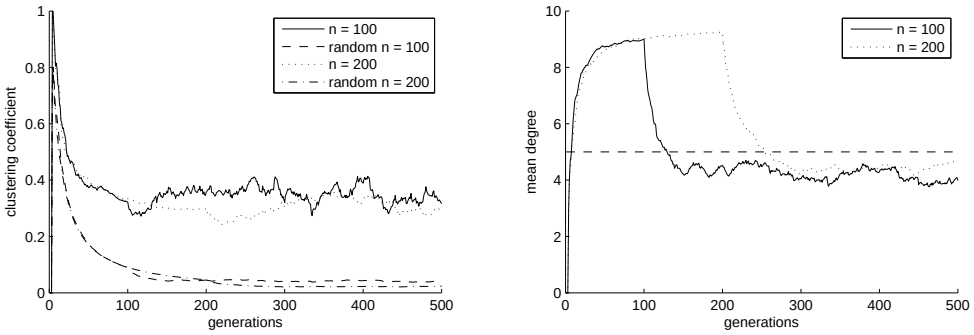

Figure 1: Development of the correlation coefficient C and mean degree $\langle k \rangle$ for two instantiations of the network model presented in section 2.2. Network parameters are $m = 5$ as well as $n = 100$ and $n = 200$ respectively. The baseline of the correlation coefficient is that of a random graph with the same network size and mean degree, $C_{random} \approx \frac{\langle k \rangle}{n}$ [Watts and Strogatz, 1998].

- population size ($n = 2, 10, 50, 100, 200$)
- population structure (fully connected or small world network, only fully connected network for $n = 2$)
- prior bias for hypothesis h_0 ($p(h_0) = 0.5, 0.6, 0.8$)
- hypothesis induction strategy (Sampling from the Posterior and Maximum a Posteriori (MAP))
- initial population configuration ($40\%, 50\%, 60\%$ of initial population having hypothesis h_1)

Every possible combination of these parameter settings was run for 50 trials to get a representative sample of the population's behaviour. Every trial run consisted of the following steps:

- initialise a network of size n, with the respective initial proportion of agents having hypothesis h_1, all others hypothesis h_0 (the distribution of hypotheses was randomised so that the order of replacement of the initial hypotheses would be different on every trial)
- for 6.000 complete generation turnovers
  - remove the oldest node from the network
  - insert a new node (with initial connections according to the respective network structure)
  - sample b datapoints from the new node's neighbours, according to their respective hypothesis

– apply the Bayesian learning algorithm described in Section 2.1 to infer a hypothesis for the new node based on the sampled datapoints

The duration of a complete generation turnover would depend on the size of the population, with 12.000 iterations for a population size of 2, and 600.000 iterations for a population size of 100, every iteration including the replacement of a node, sampling data from its neighbours and inferring an appropriate hypothesis. The bottleneck size b described in Section 2.1 was set to 3 during all simulations. The data on which the analysis is based are the time-averaged proportions of agents with hypothesis h_0 or h_1 from generation 1 until the point of observation, which were calculated and stored every 1000 generation turnovers, resulting in 6 snapshots of the development of a potential stationary distribution.

The following section describes the results and analysis of the total number of 162 simulation settings.

3 Results & Conclusions

This section discusses the results and analysis of the simulations outlined in the previous section. Details about the data analysis process and visual representation are explained in 3.1. Summaries of the behaviour observed for Sampling from the Posterior and MAP are presented in subsection 3.2 and 3.3 respectively. Subsequently, interpretations and explanations for the observed population dynamics are given: subsection 3.4 presents an in-depth description of the population dynamics, while in subsection 3.5 possible reasons for this behaviour are explored in the nature of the learning algorithm and shortcomings of the approach are pointed out. The issues highlighted there lead directly over to a general discussion of the notion of "rationality" in language acquisition in section 3.6, followed by final remarks on the learning algorithm in section 3.7.

3.1 Data Representation

The data obtained through numerical simulations as presented here is necessarily of a very different nature than the analytical results examined in previous treatments of the Bayesian Iterated Learning Model. Since the simulation results are arguably of a richer quality, not at least due to the introduction of populations, it is necessary to contemplate which features of the simulation runs to analyse.

The "baseline" against which the results can be evaluated are the analytical solutions of Markov chains, which represent the probability of an agent to have a certain hypothesis at any given point in time in a linear transmission chain. First of all, this measure has to be adopted to fit a population setting: since an entire population is very unlikely to share a unique hypothesis, the state of the system can only be described as a distribution of agents over hypotheses or, for the simple two hypotheses case examined here, the proportion of

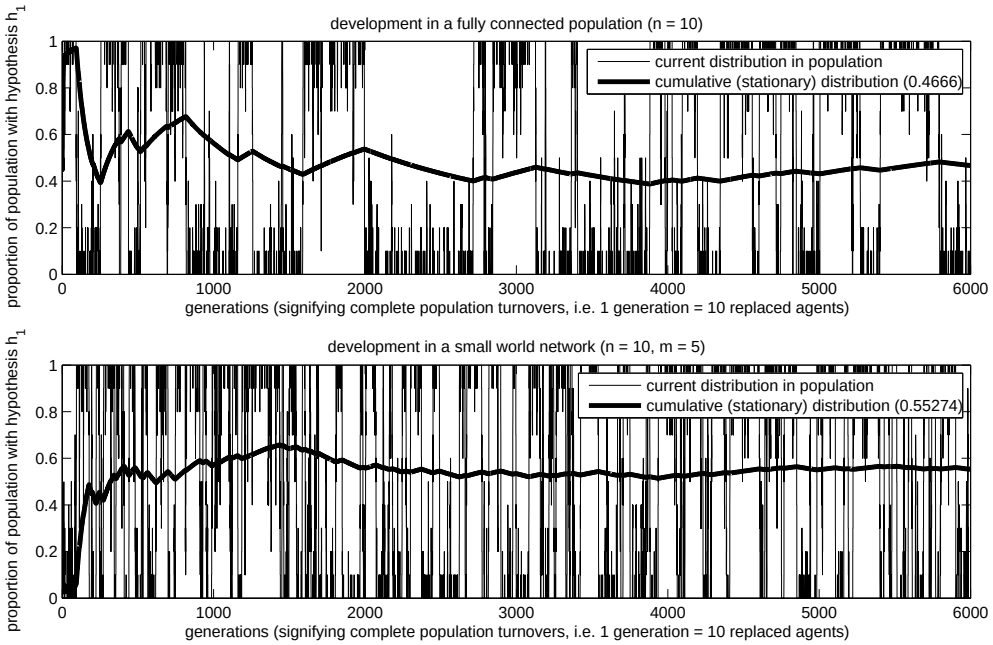

Figure 2: Single population runs in a fully connected (top) and small world network (bottom) with $p(h_0) = 0.5$, $b = 3$, $n = 10$, an initial distribution of 50% of agents with hypothesis h_1 and Sampling from the Posterior. The thick line shows the time average from the start of the simulation, which over time approximates the stationary distribution. Significantly more transitions can be observed in the small world network.

agents having a certain hypothesis[1]. This proportion is equivalent to the probability of an agent having that hypothesis at that specific point, and has to be averaged over time in order to derive the value which should over time approximate the stationary distribution.

The trace of a single simulation run can be seen in Figure 2: while the distribution of hypotheses over the population fluctuates steadily, the cumulative average appears to set to a certain value quickly. This observation is of course not a formal proof of convergence, and for many parameter settings convergence will turn out to be a lot slower than what can be observed in these simulations. The necessity of getting an intuition about the rate of convergence is the reason for the *snapshots* described in section 2.3: the cumulative averages taken every 1000 generations can be used to examine the emergence of global patterns from early fluctuations in the system. The emergence of static patterns observed in the first 6000 generations is of course not a proof of convergence either and is only used to get an intuition about the dynamics behind the individual trial runs.

Let us now turn to the type of data visualisation used throughout the rest of this work, the first example of which can be seen in Figure 3. One figure, which summarises the results for one specific prior distribution in combination with a fixed initial proportion of h_1 in the population, is comprised of 7 subfigures corresponding to different population settings: the

[1]For the rest of this work, the value denoting a system state represents the proportion of agents with hypothesis h_1.

upper row consists of the results for fully connected networks, the lower row for small world networks, both with an increasing population size n specified above the figures.

The histograms themselves show the distribution of time-averaged proportions of agents with hypothesis h_1 in the population over the 50 trials that were run for each population setting. This kind of visualisation was chosen since it enables a clear representation of global patterns across trial runs while still allowing for some insights into the development of single trials through potential fluctuations over the 6 time points from which data is presented. Another noteworthy feature is the mean value specified on top of each histogram: it represents the average probability of an agent having hypothesis h_1 *over all trials*. It is thus the entirety of the population dynamics condensed into one number, and it is this number which the analytically derived stationary distributions can be matched against.

As is obvious from Figure 3, this approximation of the stationary distribution is completely ignorant of the diversity of the population dynamics: while the value clearly converges towards 0.5 for all population structures (which in this case represents the convergence to the prior distribution), the development of each individual trial looks remarkably different, depending mainly on the population size. For small n individual populations seem to be as heterogeneous as the average over all trials suggests, while for large n populations remain homogeneous and strictly stick to either hypothesis: here the average probability of 0.5 is merely an artefact of the averaging, and a node's fate appears to be largely predetermined by the current state of the population that it is "born" into.

A more detailed discussion of the dynamics will be presented in section 3.5, but keeping the remarks presented here in mind, an analysis and interpretation of the simulation results can be attempted.

3.2 Results: Sampling from the Posterior

The "baseline" behaviour of the Sampling from the Posterior strategy has been studied most extensively and predicts convergence to the prior for linear transmission chains as well as a sensitivity to the initial population setting for a population size of 2 [Smith, 2009].

For small population sizes, particularly $n = 2$, every single population converges to the prior for all combinations of prior distributions and initial settings examined. The behaviour becomes much more complex with increasing population size, and a bifurcation seems to occur for fully connected networks: the larger the population size, the less and less dynamic single systems become, largely maintaining fully homogeneous hypotheses distributions throughout their development. Across trials however, different hypotheses win the upper hand, showing the previously observed sensitivity to initial settings: while an unbiased initial distribution (half of the agents having each hypothesis) results in an equal split between hypotheses (Figure 3), an initial proportion of 60% results in a likelihood of over 80% to converge onto that hypothesis, at least in the case of an unbiased prior (Figure 4). The average probability to obtain each hypothesis is thus also highly skewed in accordance to the initial settings.

The transition between these two extreme behaviours is also observable: the prior-

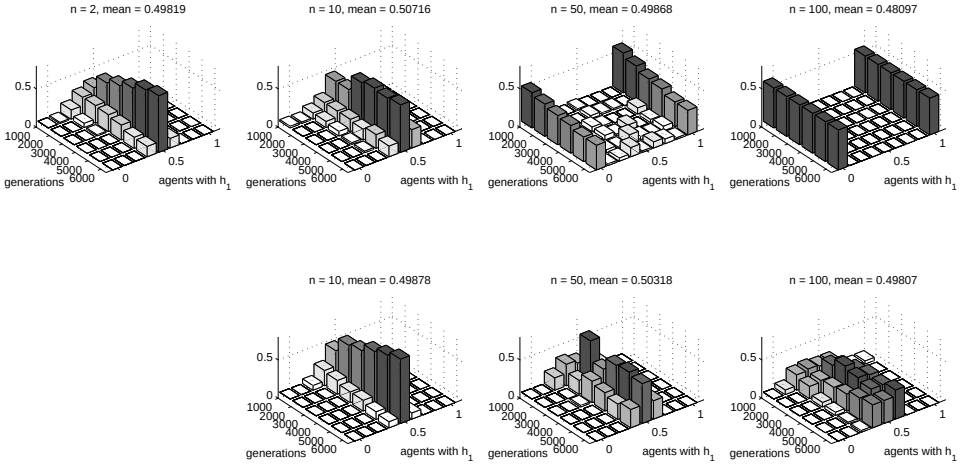

Figure 3: Distribution with prior $p(h_0) = .5$, initial proportion of 50% with hypothesis h_1 and Sampling from the Posterior for various population sizes n, fully connected networks in the top, small world networks in the bottom row. For a detailed description of the data representation please refer to Section 3.1.

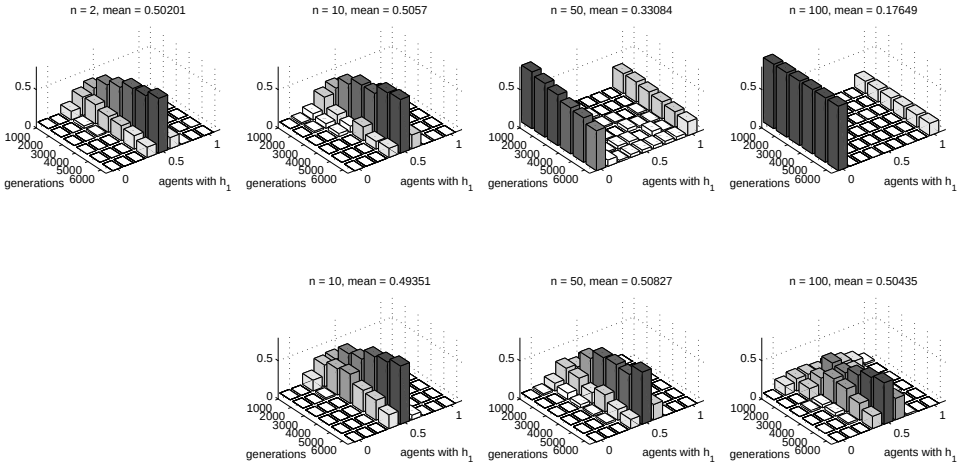

Figure 4: Distribution with prior $p(h_0) = .5$, initial proportion of 40% with hypothesis h_1 and Sampling from the Posterior for various population sizes n, fully connected networks in the top, small world networks in the bottom row. For a detailed description of the data representation please refer to Section 3.1.

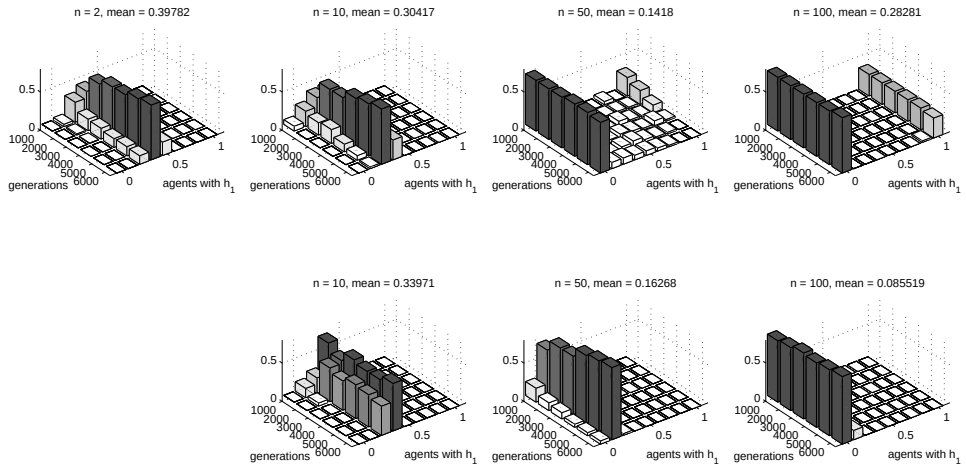

Figure 5: Distribution with prior $p(h_0) = .6$, initial proportion of 50% with hypothesis h_1 and Sampling from the Posterior for various population sizes n, fully connected networks in the top, small world networks in the bottom row. For a detailed description of the data representation please refer to Section 3.1.

determined peaked distribution for $n = 2$ starts to get broader and ultimately flattens out with increasing population size, reflecting more variability between different trial runs. Ultimately, all trials stick to one hypothesis only, resulting in the distinct split between the two homogeneous situations.

The same behaviour is observed for the development in small world networks. Here, however, the populations retain their dynamic character for much larger population sizes – the distributed character of the networks seems to enhance the avoidance of a completely static system state.

Interactions between parameters become more complex for biased prior distributions, as can be seen in Figures 5 and 6. While the populations which are unaffected by initial settings still converge to the prior, the diversification and broadening of the distribution of averages now exhibits a distinct shift towards the direction of the bias in fully connected networks. Most notably, the distribution does not appear to *flatten out* in between, with the peak instead simply shifting towards the preferred hypothesis. For even bigger populations the split between the strictly homogeneous populations can be observed again, only that this time the distribution of populations appears to be modulated by the prior: while a bias in favour of a hypothesis helps to draw a bigger number of populations towards homogeneous acceptance of that hypothesis, the outcomes are still mostly determined by the initial settings of the population.

In small world networks different initial settings appear to affect only the early development of the hypotheses distribution: the shift towards the preferred hypothesis for intermediate population sizes resembles that of the fully connected network, but the completely

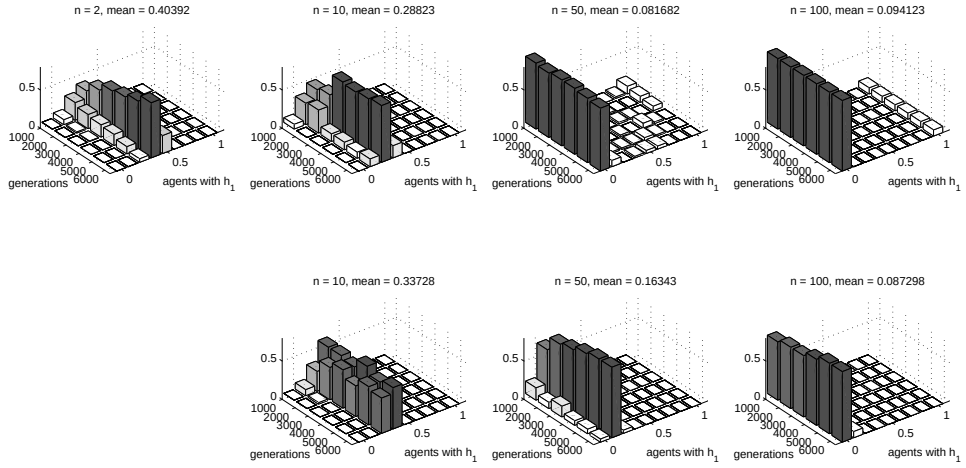

Figure 6: Distribution with prior $p(h_0) = .6$, initial proportion of 40% with hypothesis h_1 and Sampling from the Posterior for various population sizes n, fully connected networks in the top, small world networks in the bottom row. For a detailed description of the data representation please refer to Section 3.1.

stationary state of the largest networks is, just like the behaviour with unbiased priors, dependent on the prior biases alone. The biases are in fact amplified so that *all* simulation runs quickly converge onto the favoured hypothesis, with only slight but continuous fluctuations away from it.

The behaviour described thus far becomes even more extreme with stronger biases. Here, all but the smallest populations converge to homogeneous acceptance of the bias-preferred hypothesis, with a bias of .8 sufficing to drag every single trial to a single-hypothesis sink for all tested parameter settings. Only the smallest fully connected population with two agents allows for more variation, with every trial exhibiting convergence to the prior.

3.3 Results: MAP

The populations' development using *Maximum a Posteriori* exhibits the same two behavioural extremes described in the previous subsection, albeit with different convergence results and different onsets for the distinct system behaviours. As can be seen for the larger populations in Figure 7, MAP appears to be more sensitive to the prior distributions than Sampling from the Posterior (cf. Figure 4), a property which is in accordance with similar behaviour in linear transmission chains, where MAP would also exaggerate the effect of the prior [Kirby et al., 2007, Griffiths and Kalish, 2007].

The properties of the more dynamic system behaviours for smaller populations are remarkably different: while a peaked distribution can be observed again, it always converges to .5, *independent* of the prior biases examined here (cf. Figures 8 and 9). This odd behaviour is indeed caused by sampling effects: with a bottleneck size of 3 items and an error term of

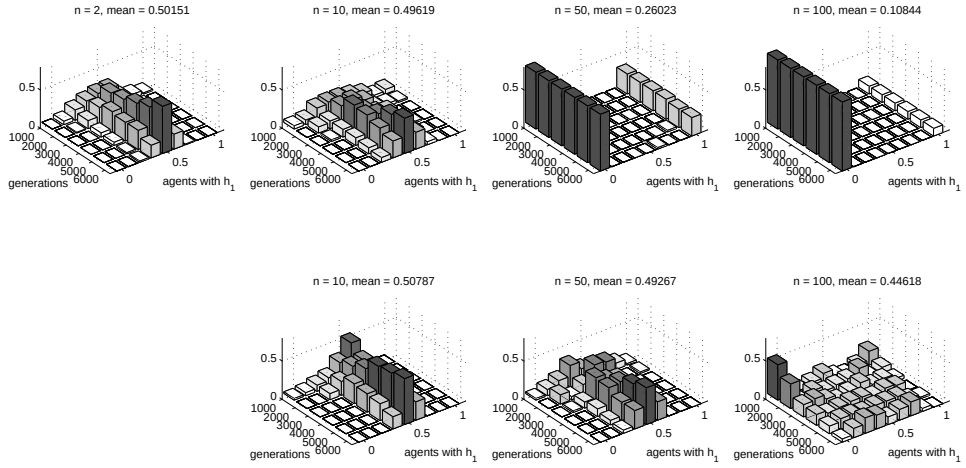

Figure 7: Distribution with prior $p(h_0) = .5$, initial proportion of 40% with hypothesis h_1 and MAP for various population sizes n, fully connected networks in the top, small world networks in the bottom row. For a detailed description of the data representation please refer to Section 3.1.

.05, the border case (2 items corresponding to one, the third item corresponding to the other hypothesis) is not sensitive to multiplying the probability of the data with biases as extreme as .8 and .2 respectively, i.e. $\arg\max_h p(d|h) = \arg\max_h p(d|h)p(h)$ for this combination of parameters.

While the effects of the prior biases can thus only be observed with wider bottleneck sizes or an adjusted error term, the behaviour of the MAP strategy in a population with an unbiased prior can still be studied: the flattening out of the peak appears to happen much more slowly, while the occurrence of the extreme, static system behaviour in fully connected networks seems to be more abrupt.

The question of whether there is a continuous transition between these two behaviours as was observed with Sampling from the Posterior cannot be answered from the data that was obtained, but assuming a similar relationship between the behaviour of the fully connected and small world network populations as observed using the Sampling strategy, the following situation might be deduced: for the parameter settings under examination, which are identical to those used with the Sampling condition, the small world networks do not exhibit the extreme, static system state at all. Besides showing no detectable effect of different initial settings, large populations underly heavy fluctuations throughout the full observed period of 6000 generation turnovers, with every single trial exhibiting only a very weak trend towards the .5 mark. Since the average across trials is still very clearly around .5, it is possible that the increased population size is only delaying the formation of such an easily detectable distribution. This suggestion is supported by the decreasing clearness of the peaked distribution for smaller population sizes which moreover also exhibit significant fluctuations among single trials. This transition might indicate that a distinct peaked distribution would be observable

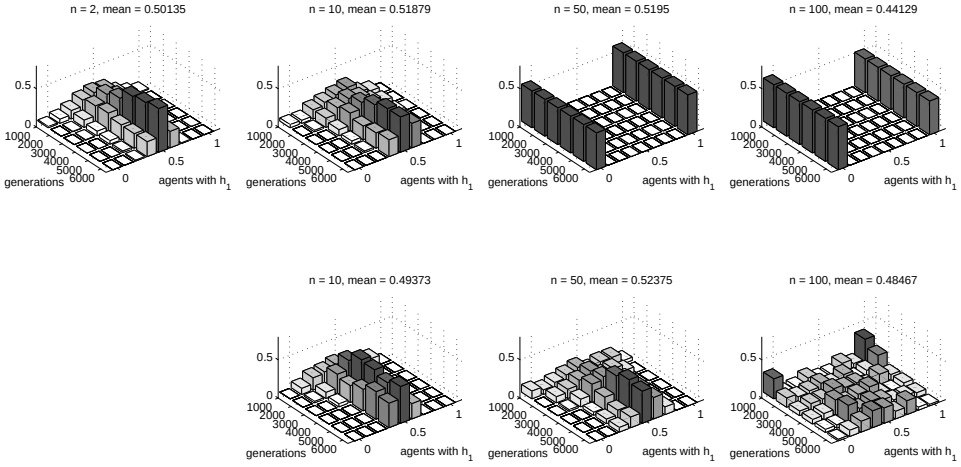

Figure 8: Distribution with prior $p(h_0) = .5$, initial proportion of 50% with hypothesis h_1 and MAP for various population sizes n, fully connected networks in the top, small world networks in the bottom row. For a detailed description of the data representation please refer to Section 3.1.

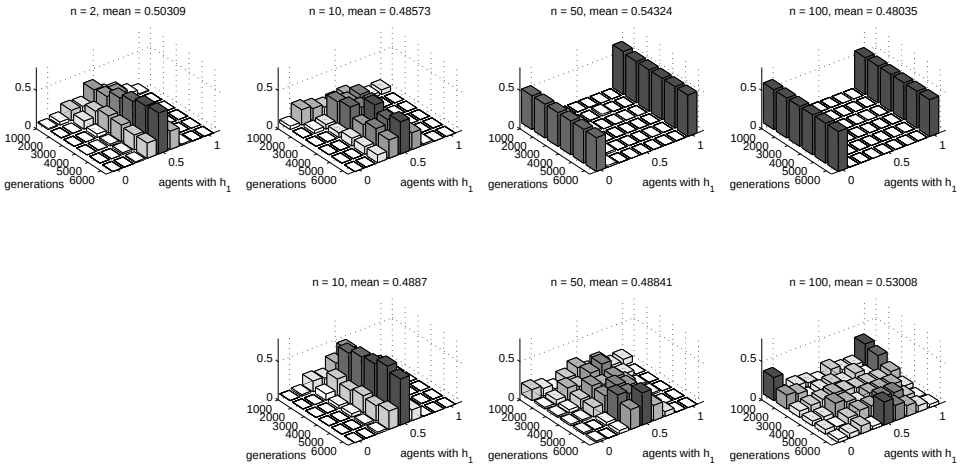

Figure 9: Distribution with prior $p(h_0) = .6$, initial proportion of 50% with hypothesis h_1 and MAP for various population sizes n, fully connected networks in the top, small world networks in the bottom row. For a detailed description of the data representation please refer to Section 3.1.

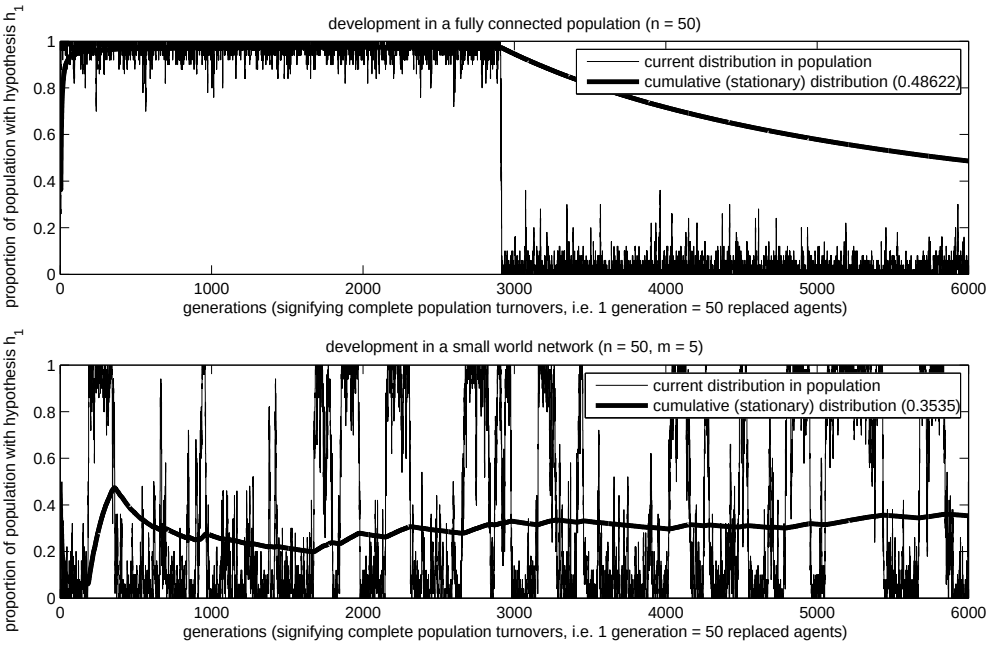

Figure 10: Single population runs in a fully connected (top) and small world network (bottom) with $p(h_0) = 0.5$, $b = 3$, $n = 50$, an initial distribution of 50% of agents with hypothesis h_1 and Sampling from the Posterior. The thick line shows the time average from the start of the simulation, which over time approximates the stationary distribution. Note the significantly reduced number of transitions in comparison to the smaller population seen in Figure 2.

given longer observation windows, or otherwise by simply increasing the population size.

3.4 Population Dynamics

Before we attempt to draw general conclusions about the different behaviours documented in this section, a more precise study of the underlying source of the distributions – the developmental dynamics of single simulation trials which were already hinted at – seems to be appropriate.

One such trial along with the development of its average was already depicted in Figure 2. While a trial run using the same parameter settings in a larger population as seen in Figure 10 is of a qualitatively similar nature, the number of transitions between the two extremes decreases significantly. Further increases in population size bring the dynamics to a standstill: once caught in either homogeneous extreme, populations are less and less likely to escape this situation. Both of the competing hypotheses, irrespective of their prior probabilities, appear to act as *sinks* with only slight variations caused by noise through mislearning (in the case of Sampling from the Posterior) or the correct learning of "deviant" samples from the current population. The force of each attractor can only be overcome by cumulative noise, shifting the proportion in the population over the 50% mark and thus simply pushing the population into the region of the other attractor, ultimately resulting in convergence on the complementary hypothesis.

It might be asked where this tendency to comply with other individuals in the population, which could be interpreted as some kind of functional, communicative pressure, comes from when there is no selection for communicative efficiency or similar built into the model.

When the population size is increased, the functional pressure to adhere to any established *norm* becomes stronger: while in a linear chain a single *mislearning* or the sampling of a few unprototypical datapoints sufficed to cause a complete population "turnover", the introduction of more agents which will potentially keep the majority hypothesis drastically reduces the likelihood of a changeover taking place. This regulatory pressure produced by the interactions between the agents thus introduces a kind of inertia into the system which was not present in linear transmission chain analyses, which would predict a much more diverse and vivid development. Looking at it from another perspective, learners embedded in a population are to a large extent slaves to the input they receive: for the case of language, the development towards grammars which might be favoured in terms of learnability is ultimately limited by the linguistic input from the environment, a point which will be made more clear in the next section.

As can be extrapolated from the results, this normalising behaviour becomes stronger and stronger when increasing both population size n as well as bottleneck size b, since bigger data samples are more representative and lead to less mislearning. Crucially, both these parameter changes would appear plausible regarding the modelling of a real world setting.

The sensitivity to the initial frequencies of the different hypotheses in the population is simply a consequence of the population pressure described in the previous paragraph. For large population sizes which exhibit static behaviour the initial heterogeneous distributions are inherently unstable and quickly collapse onto either homogeneous setting, again mainly guided by stochastic processes, i.e. random walks.

The slight modulations of the distribution of convergence outcomes caused by different bias settings sheds a light on the effects of biases within these larger populations: while an equal initial proportion of both hypotheses with unbiased priors leads to a 50:50 split (see Figure 3), it is not only the initial distribution which can affect the probability of converging to either hypothesis (see Figure 4), but also the biasing through the priors: as can be seen in Figure 5, a biased prior can influence the direction of convergence for individual trials towards the direction of the preferred hypothesis. Interactions become more complex when initial distributions as well as priors are biased (e.g. Figure 6), with the prior appearing to be the weaker force. This bias *pull* is most obvious in pre-convergence population settings, in interaction with the effects of random noise. Its effect on the likelihood or direction of population turnovers once a population has converged could not be examined at this point, but due to its apparent coupling to the random walk processes which become less significant in bigger populations, it is also likely to play an even smaller role in these cases, at least under moderate bias settings.

It is important to point out that population turnovers are of course not strictly impossible – they only become less and less likely with increasing population size. Since the amount of cumulative noise required to overcome the dominating hypothesis of the current popula-

tion becomes larger and larger, the turnovers are pushed towards the limits of observability within numerical simulations. This suggests that the results obtained for large population sizes might in fact not be approximations of a stationary distribution, but highly skewed intermediate results before the occurrence of a very unlikely population turnover. Given that the homogeneous period following such a turnover might last as long as the previous one, a longer observation of these populations would also result in peaked distributions similar to those obtained for smaller population sizes, and indeed also linear transmission chains. Consequently, the results presented here do not invalidate the results obtained from analytical treatments, rather than emphasise the role of large populations as a conservating factor which might seriously slow down the rate of convergence, resulting in significantly lower convergence rates than previously thought [Rafferty et al., 2009].

Another phenomenon which demands explanation are the apparent different developmental speeds between the two population structures, fully connected and small world networks. While in a fully connected population the *communicative pressure* or *learning dependence* can potentially be forced upon an agent from every single individual in the entire population (particularly given a large bottleneck size), the local dense clusters of small world networks allow these pressures to be temporarily overcome locally (see Figure 11). Thus the amount of "noise" over a majority variant is amplified by networks with the small world property which also makes entire population shifts more probable in these kinds of networks. Following the reasoning of the previous paragraph, while small world networks do not make a *major* difference in terms of population dynamics, they do not slow down convergence rates as much as fully connected networks do.

The particular network model used here exhibits constant fluctuations in its more global network and community structure: given more constant community patterns within the small world networks they would not only allow for more "individualism" which might cumulatively result in wider spread of a minority variant as observed in the simulations, but might also exhibit longer steady periods of minority-variant usage within certain parts of the population. This again would not only increase the rate of convergence, but might also lead to less extreme synchronic hypothesis distributions among the population.

3.5 Bayesian Rationality and Language Change

Given this analysis of the population-level behaviour of the Bayesian Iterated Learning Model, it would be an interesting task to investigate the plausibility of the observed dynamics. The expected behaviour in terms of phase transitions in *language evolution*, the question on how non-strict language universals might emerge from a population, can only be a matter of speculation, since such behaviour has never been observed. However, a wealth of qualitative descriptions on the question of *language change*, as discussed earlier, constitutes a potential baseline against which the dynamics of the model can be compared.

As might already have become clear, the current design of the model is not unifiable with properties normally associated with language change at all. Since both hypotheses constitute

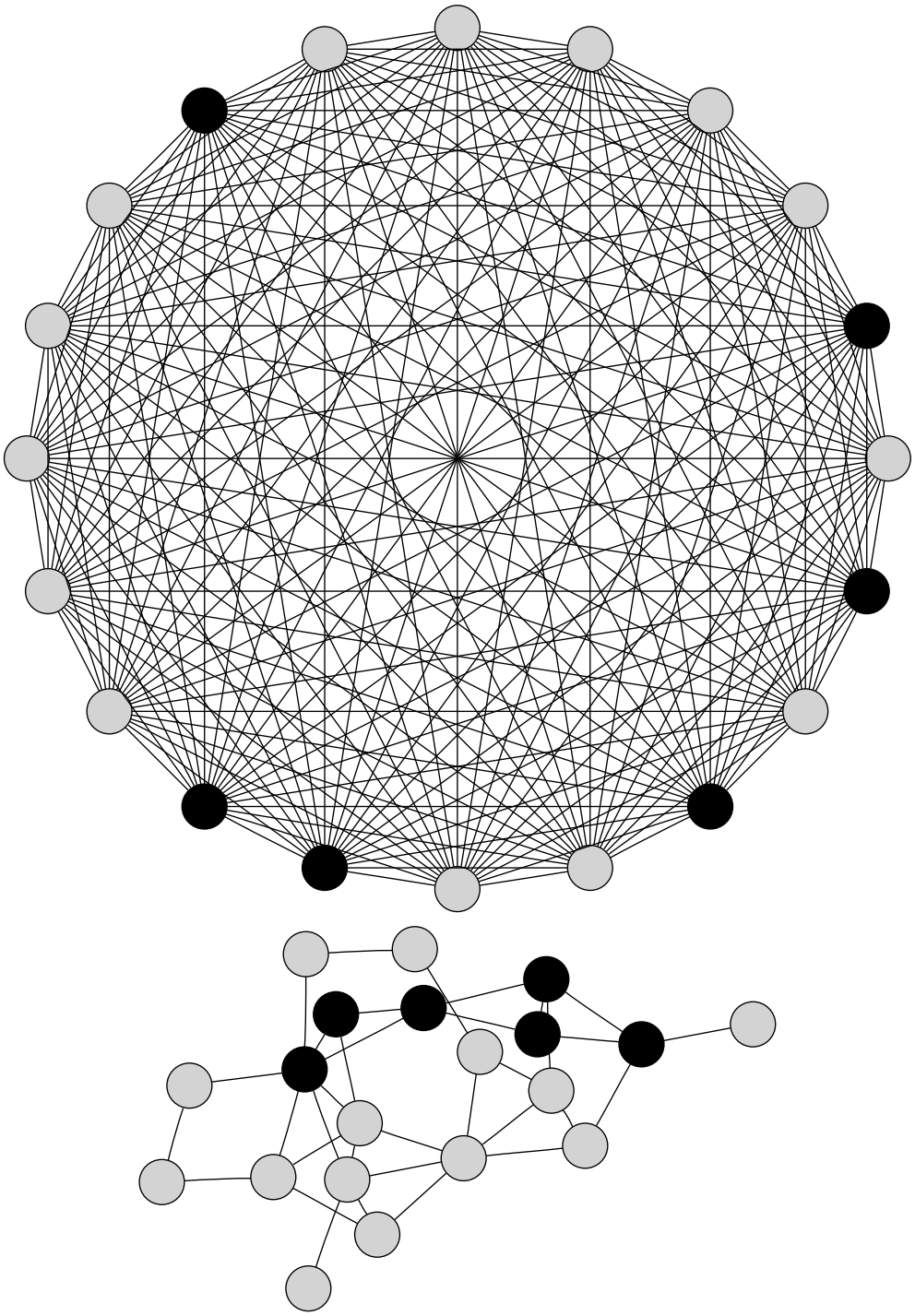

Figure 11: Visualisation of the distribution of variants h_0 (grey) and h_1 (black) during simulation runs in populations of 20 agents in a fully connected (top) and small world network (bottom). While the proportion of agents with the minority variant h_1 is equal in both populations, the adoption of this variant arises out of local clustering effects in the small world network, whereas the emergence of the variant in the regular network is random.

very strong sinks which are amplified by bigger populations and wider bottlenecks which help to keep the currently established variants in the majority, no logistic patterns of change can be observed between them. Instead of a low threshold from which change might pick up and spread, there is a clear bifurcation at 50% from which all changes tend to develop towards either extreme.

This deficiency in terms of regarding the model as a model of language change, its inability to cope with *phase transitions* has already been critisised [Niyogi and Berwick, 2009], and is somewhat *by design*. The basic flaw found in this simplest model of Bayesian Iterated Learning is that the process of "mislearning" a hypothesis is possible in both directions, a feature that is usually not found in processes of language change. Real world examples for this might be plosives that become affricated while affricates do not tend to lose their affrication, or *V2* grammars collapsing onto *SVO* but not the other way around.

In previous models of language change discussed in the literature review, this issue of unidirectionality of change was often circumvented by implementing a strong bias for the new variant, often accompanied by the inability to return to the old variant. This is a simple yet legitimate strategy since it is the functional advantage of a variant over another which is responsible for the increased use of new variants in the first place, and also the resulting S-shaped pattern which the simple Bayesian model studied here fails to account for.

A weakness of the language change models presented earlier is that they are usually only designed to exhibit a wave of change once, often modelled to reproduce certain documented periods of change throughout human history. What would be desirable is a more dynamic model of language change without the inclusion of sinks. The simplest model of this kind would thus have to consist of three distinct variants where change can only occur in one direction, exhibiting circular patterns of change. The strength of such a toy model would be the possibility to study the effects of different network types and finer topological properties on the triggering and overall frequency and speed of changes taking place.

The question is how the Bayesian model has to be adopted to accommodate such circular change patterns, or strictly unidirectional changes in the first place. While the latter could be achieved by extremely biased priors, this would also establish the preferred hypothesis as a sink and would thus make a transfer over to the next hypothesis in the circle unlikely. Consequently, it is the structure of the different emission probabilities for each hypothesis that has to be adopted to allow unidirectional change. Here, the learner's assumed perfect knowledge of the hypotheses causes a kind of dilemma of rationality within the Bayesian approach: disregarding situations of extreme noise where change would mainly be governed by sampling effects, in order for data emitted based on a certain hypothesis A to be reinterpreted as coming from another hypothesis B, the emission probabilities of B have to be adapted to account for a good amount of the data from A, otherwise a learner is very unlikely to end up inferring hypothesis B from the input. But since the emission probabilities used to induce the hypothesis reflect the actual probabilities with which different outputs are produced (assuming that the hypotheses structures are identical across agents), this means that the adaptation made to accommodate change in one direction results in an increased

emission of utterances prototypical for the previous hypothesis, again making a mislearning taking place in the other direction more likely. The difficulty of achieving unidirectional change with a rational reasoner poses a significant problem to the dynamics of the BILM, and the author of this work is unable to see potential solutions to this issue. The necessity of having to account for every observed input without the possibility to discard or ignore single datapoints in the inference process appears to be a weakness of the framework. It also seems like the introduction of the error term was motivated by the model's inability to cope with variation in the first place. But by making this term, and consequently also the production and accomodation of glitches or *misproductions*, an integral part of the hypothesis structures (at least in this simplest model setting), accounting for such glitches effectively becomes part of the language model.

The issue pointed out here might seem to be just a matter of implementation, but the argument can also be made on a more theoretical level, which suggests an incompatibility between the assumptions of a rational learner and the processes underlying language change: as was discussed earlier, change is triggered and propagated by the *reinterpretation* or *reanalysis* of linguistic input that a learner receives.It is precisely this process which appears to be irreconcilable with the core assumptions of the Bayesian learning framework: given the observed data, a Bayesian learner uses its knowledge of the emission probabilities to infer the hypothesis underlying the observations. A hypothesis which is not the one actually underlying the data he or she received would only be chosen given very strong sampling effects resulting in a hugely different distribution of the input data. In some sense, assuming a rational learner appears to essentially rule out the very process on which language change is based.

As a consequence it could be argued that, in order to accommodate language change, language learners do indeed have to be *naive* when it comes to the learning task, and *must not* try to second-guess the *intentions* or underlying production mechanisms of their interlocutors, which is how the Bayesian approach could be characterised. The ease with which unidirectional, logistic change patterns are produced using cue-based language acquisition models further questions the usefulness of rational models of language learning. Moreover, it is hard to see how rational inference might be extended to a realistic language learning task, since the assumptions regarding perfect knowledge of a much larger parameter or grammar space become highly unrealistic.

The issues pointed out here call into question the utility of the Bayesian framework as a model of language change and all aspects of language that undergo "classical" processes of change. Whether the argument can be extended to the model's status as a tool to investigate cultural language evolution depends on the assumptions about the nature of the processes underlying this evolution: supposing that emergent universal properties of language evolve through similar processes as those observable for "simple" language change, a thorough revision of the applicability of the BILM seems to be appropriate. If one does not wish to impose constraints on the dynamics by which such properties develop, and one accepts a model which relies on random fluctuation alone as a satisfactory explanation for the phenomena at hand,

then there is no problem with the use of the BILM.

Admittedly, some of the criticism voiced in this section is rather harsh, particularly regarding the absence of concrete claims concerning the applicability of the Bayesian framework, and particularly the simple models examined so far, as realistic models of language acquisition. In the following section, potential strategies undertaken to counter the issues highlighted so far will be discussed.

3.6 Bayesian Rationality in Mixed Heterogeneous Populations

Putting aside the criticisms voiced thus far, there are other deficiencies that a Bayesian model of language acquisition might have to address in order to become more realistic. As was previously mentioned, the promise of rational inference relies on the accurate knowledge of the actual probabilities with which signals are to be observed, or at least a model thereof. It was also noted that these requirements are strictly speaking not met for usage in a population such as used in the simulations here, and this section aims to address this issue briefly. The ways in which Bayesian inference might be designed when learning from a finite, necessarily heterogeneous population in the first place are manifold, and the potentially wildly different outcomes have been noted [Ferdinand and Zuidema, 2009]. The possibility to incorporate knowledge of the different identities of multiple input sources, which is ongoing work, would allow for much more refined ways of inference. Instead of the trivial data-based induction of the most likely hypothesis as in the current model, the increase in explanatory power would allow the introduction of arbitrary speaker-based utility functions over competing variants [Zuidema, personal communication] which might remedy the insufficiencies of the reported dynamics. Moreover, such an approach would allow for an effective treatment of multilingualism or selective usage of variants in different contexts, a particularly hard problem for any model of language acquisition or production.

The possibility of handling such complex issues quite easily highlights the power of the Bayesian framework for the modelling of complex abstract reasoning processes but, in concordance with the thoughts presented in the previous section, it again raises the question of how much abstract *thinking* or *inference* should be required by a model of language acquisition in the first place, when models based on naive data-centred approaches like trigger-based learning exist and prove to be more explanatorily adequate.

In a very interesting fashion these two extreme positions regarding the *intentions* underlying the process of language acquisition recapitulate an ongoing discussion in the field of historical linguistics on the role of identity in language change, particularly dialect formation: based on extensive analysis of the outcome of dialect mixture following the colonalisation of New Zealand by speakers of various different Englishes [Trudgill et al., 2000], a formal account of processes involved in dialect mixture has been proposed [Trudgill, 2004]. Crucially, it claims that the outcome of the subsequent dialect formation can be deducted in a mostly deterministic fashion based on the distribution and frequency of variants across the original language substrate. This "mechanical view" means to explain the resulting dialectal prop-

erties through simple reproduction mechanisms [Trudgill, 2008a], i.e. language acquisition in children [Trudgill, 2008b] and mutual *accommodation* in adults. This deviates from most sociolinguistic treatments which would often seek explanations for specific aspects of dialect formation in the question of identity speakers or communities [Holmes and Kerswill, 2008, Schneider, 2008]. The theory is still heavily disputed and scrutinised, also on the basis of computational modelling [Baxter et al., 2009]. Should the strictly mechanical view turn out to be a useful general assumption regarding questions of identity in dialect mixing in adults, then it is even more questionable how they might play a role for language acquisition in infants, thus also supporting less powerful models of language learning.

3.7 Other Issues

While the most fundamental weaknesses of the Bayesian Iterated Learning Model have been discussed in detail, there are two more issues which should be considered for the construction of future models.

On a more general note, the hypothesis structure of the model examined here only allows the strict acceptance of one hypothesis and lacks the ability to account for systematic variation within single agents. This issue has been addressed in an enhanced Bayesian model using Beta distributions to model gradual belief into more than one hypothesis [Reali and Griffiths, 2008], and it has also been tested in experimental settings [Reali and Griffiths, 2009]. Preliminary experiments using this continuous variable model for language acquisition in a population did not yield convergence onto either homogeneous grammar (which was the result of previous analytical as well as experimental investigations of the model) unless unrealistically high biases for overgeneralisation in the region of $\alpha < .0001$ were used, and were consequently not pursued further.

Another issue lies in the choice of the inference strategy employed: while Sampling from the Posterior is far more well-studied and has also been shown to produce better fits for experimental data [Reali and Griffiths, 2009], its indeterministic nature does not seem psychologically plausible, at least for the case of language. The robustness of human language acquisition has been emphasised before and should be reflected in the nature of any computational acquisition algorithm, suggesting that MAP should be preferred over the Sampling strategy.

4 Summary & Future Work

In this work, two distinct strands of enquiry were pursued: the extension of the Bayesian Iterated Learning Model to a population setting on one, and the investigation of the effects of small world networks on the behaviour of populations engaging in a language acquisition task on the other hand. While both issues were intimately linked throughout the treatment, it is worth summarising the conclusions about each separately.

For the first time, the system dynamics of the simple Bayesian Iterated Learning Model

used in this work were examined numerically in a population setting. The model was shown to exhibit only random fluctuations which were largely suppressed by the majority variant used by the community given a large enough population size. The observed population dynamics have two major consequences for the interpretation of previous results within the BILM: firstly, the largely static system behaviour which emerged with increasing population size constitutes a significant factor which has been left of the study of convergence times so far. Both larger populations and a larger bottleneck size which are realistic assumptions regarding a real world setting would significantly slow down the population dynamics. Secondly, it suggests that the diachronic convergence on a stationary distribution is not reflected in the synchronic composition of speaker communities, for which the model predicts mostly homogeneous distributions due to communicative or learning constraints between the agents.

A number of shortcomings of the simple Bayesian model as a model of *language change* rather than of cultural evolution in general have been pointed out. The BILM failed to account for systematic mislearning in language acquisition and it was shown how the assumption of perfect knowledge of the underlying hypotheses might be a significant obstacle in reconciling a Bayesian model of language acquisition with the concept of language change. This flaw was also reflected in the population dynamics, which did not resemble dynamics documented for cases of language change at all. The fact that population dynamics have been largely disregarded in previous treatments of the BILM goes in line with a number of other analytical models of cultural transmission which are not explicit in how exactly their results can be related to real world processes. While models particularly dedicated to language change have real world data which they can be evaluated against, models of cultural evolution seem to evade this kind of scrutiny, leaving the exact relation between model results and reality open to interpretation. Since the dynamics and even the setup of the BILM were shown to be unrealistic in terms of modelling language change, it is not straightforward to relate the results of such a toy language model to the effects of biases on real languages. Many models of cultural transmission have not been explicit about the distinction between cultural evolution on one and concrete investigations of language change on the other hand, even though the two approaches differ significantly in terms of their direct applicability. This distinction should be made more explicit in future treatments and a more paradigmatic set of constraints on possible dynamics of cultural transmission processes should be established. Crucially, if this set of constraints differs from those observed for language change, then more caution should be taken when relating results on the evolution of other cultural traits to language: while many such traits are not only learned but also *taught*, language is acquired largely by imitation but without explicit teaching, and thus potentially also underlies different processes of *change*.

The second major contribution of this work was to study the effects of population structure on the development of cultural evolution by introducing small world networks to model richer populations dynamics. While the population dynamics did not differ qualitatively from those for a fully connected community, the distributed network structure significantly amplified diversity within the population, thus altering the transmission and development of

cultural traits. By using a more dynamic language acquisition algorithm future investigations should be able to relate the observed effects to real world phenomena more easily. The basic requirements for such a dynamic model of language change which does not exhibit a pretermined outcome like most previous investigations has been outlined. Ideally such a model would exhibit many of the features of language change that have been studied descriptively, from S-shaped curves to the triggering of language change by agents at the periphery of tight-knit communities which could be traced effectively through the usage of small world network models. Once such a model is constructed it can be used to investigate the effects of particular population structures in more depth, e.g. effects of population size, density or degree of clustering on the stability of languages as well as the likelihood and speed of changes.

While many weaknesses of the acquisition algorithm used in this work were pointed out, the ongoing development of more complex Bayesian models promises more accurate predictions of the dynamics of cultural transmission and particularly language change. Both research on such models as well as more systematic studies of the effects of population structures on cultural transmission are likely to play an important role in future investigations of cultural transmission through Iterated Learning Models.

References

[Amaral et al., 2000] Amaral, L. A. N., Scala, A., Barthélémy, M., and Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21):11149–11152.

[Barabási and Albert, 1999] Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286:509–512.

[Baronchelli et al., 2006] Baronchelli, A., Dall'Asta, L., Barrat, A., and Loreto, V. (2006). Topology induced coarsening in language games. *Physical Review E*, 73:015102.

[Baronchelli et al., 2008] Baronchelli, A., Loreto, V., and Steels, L. (2008). In-depth analysis of the naming game dynamics: the homogeneous mixing case. *International Journal of Modern Physics C*, 19(5):785–812.

[Baxter et al., 2006] Baxter, G. J., Blythe, R. A., Croft, W., and McKane, A. J. (2006). Utterance selection model of language change. *Physical Review E*, 73:046118.

[Baxter et al., 2009] Baxter, G. J., Blythe, R. A., Croft, W., and McKane, A. J. (2009). Modeling language change: An evaluation of Trudgill's theory of the emergence of New Zealand English. *Language Variation and Change*, 21:257–296.

[Blythe and Croft, 2009] Blythe, R. A. and Croft, W. A. (2009). The speech community in evolutionary language dynamics. *To appear in Language Learning*.

[Bornholdt and Ebel, 2001] Bornholdt, S. and Ebel, H. (2001). World Wide Web scaling exponent from Simon's 1955 model. *Physical Review E*, 64:035104.

[Brighton and Kirby, 2001] Brighton, H. and Kirby, S. (2001). *Lecture Notes in Computer Science*, chapter The Survival of the Smallest: Stability Conditions for the Cultural Evolution of Compositional Language, pages 592–601. Springer.

[Briscoe, 2002] Briscoe, E., editor (2002). *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge University Press.

[Briscoe, 2000] Briscoe, T. (2000). Macro and micro models of linguistic evolution. In *3rd International Conference on Language and Evolution*.

[Bynon, 1977] Bynon, T. (1977). *Historical Linguistics*. Cambridge University Press.

[Chambers and Trudgill, 1980] Chambers, J. and Trudgill, P. (1980). *Dialectology*. Cambridge University Press.

[Chomsky, 1957] Chomsky, N. (1957). *Syntactic Structures*. Mouton.

[Chomsky, 1965] Chomsky, N. (1965). Aspects of the theory of syntax. Special technical report 11, Massachusetts Institute of Technology Research Laboratory of Electronics.

[Christiansen and Kirby, 2003] Christiansen, M. H. and Kirby, S., editors (2003). *Language Evolution*. Oxford Press.

[Clark and Roberts, 1993] Clark, R. and Roberts, I. (1993). A computational model of language learnability and language change. *Linguistic Inquiry*, 24(2):299–345.

[Clark, 1997] Clark, R. A. J. (1997). Language Acquisition and Implications for Language Change: A Computational Model. In *Proceedings of the GALA 97 Conference on Language Acquisition*, pages 322–326.

[Cornish, 2006] Cornish, H. (2006). Iterated learning with human subjects: an empirical framework for the emergence and cultural transmission of language. Master's thesis, School of Philosophy, Psychology and Language Sciences, University of Edinburgh.

[Croft, 2000] Croft, W. (2000). *Language change: an evolutionary approach*. Longman.

[Dall'Asta et al., 2006a] Dall'Asta, L., Baronchelli, A., Barrat, A., and Loreto, V. (2006a). Agreement dynamics on small-world networks. *Europhysics Letters*, 73(6):969–975.

[Dall'Asta et al., 2006b] Dall'Asta, L., Baronchelli, A., Barrat, A., and Loreto, V. (2006b). Non-equilibrium dynamics of language games on complex networks. *Physical Review E*, 74(3):036105.

[Davidsen et al., 2002] Davidsen, J., Ebel, H., and Bornholdt, S. (2002). Emergence of a small world from local interactions: Modeling acquaintance networks. *Physical Review Letters*, 88:128701.

[Eckardt, 2008] Eckardt, R. (2008). *Variation, Selection, Development Probing the Evolutionary Model of Language Change*, chapter Introduction, pages 1–22. Mouton.

[Ferdinand and Zuidema, 2009] Ferdinand, V. and Zuidema, W. (2009). Thomas' theorem meets Bayes' rule: a model of the iterated learning of language. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

[Gibson and Wexler, 1994] Gibson, E. and Wexler, K. (1994). Triggers. *Linguistic Inquiry*, 25(3):407–454.

[Girvan and Newman, 2002] Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

[Gold, 1967] Gold, E. M. (1967). Language identification in the limit. *Information and Control*, 10:447–474.

[Granovetter, 1973] Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, 78(6).

[Granovetter, 1983] Granovetter, M. S. (1983). The Strength of Weak Ties: a Network Theory Revisited. *Sociological Theory*, 1:201–233.

[Griffiths and Kalish, 2005] Griffiths, T. L. and Kalish, M. L. (2005). A bayesian view of language evolution by iterated learning. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*.

[Griffiths and Kalish, 2007] Griffiths, T. L. and Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, 31:441–480.

[Holmes and Kerswill, 2008] Holmes, J. and Kerswill, P. (2008). Contact is not enough: A response to Trudgill. *Language in Society*, 37(2):273–277.

[Jin et al., 2001] Jin, E. M., Girvan, M., and Newman, M. E. J. (2001). The structure of growing social networks. *Physical Review E*, 64:046132.

[Ke et al., 2008] Ke, J., Gong, T., and Wang, W. S.-Y. (2008). Language change and social networks. *Communications in Computational Physics*, 3:935–949.

[Kerswill and Williams, 2000] Kerswill, P. and Williams, A. (2000). Creating a New Town koine: Children and language change in Milton Keynes. *Language in Society*, 29:65–115.

[Kirby, 2001] Kirby, S. (2001). Spontaneous evolution of linguistic structure – an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2):102–110.

[Kirby, 2002] Kirby, S. (2002). *Linguistic evolution through language acquisition*, chapter Learning, Bottlenecks and the Evolution of Recursive Syntax, pages 493–. In [Briscoe, 2002].

[Kirby et al., 2008] Kirby, S., Cornish, H., and Smith, K. (2008). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686.

[Kirby et al., 2007] Kirby, S., Dowman, M., and Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, 104:5241–5245.

[Kirby and Hurford, 1997] Kirby, S. and Hurford, J. R. (1997). Learning, culture and evolution in the origin of linguistic constraints. In Husbands, P. and Harvey, I., editors, *Proceedings of the Fourth European Conference on Artificial Life*, pages 493–502. MIT Press, Cambridge.

[Komarova and Nowak, 2003] Komarova, N. L. and Nowak, M. A. (2003). Language dynamics in finite populations. *J. theor. Biol.*, 221:445–457.

[Kossinets and Watts, 2006] Kossinets, G. and Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science*, 311:88–90.

[Kroch, 1989a] Kroch, A. (1989a). *Language Change and Variation*, volume 52, chapter Function and grammar in the history of English: periphrastic do, pages 133–172. John Benjamins Publishing Company.

[Kroch, 1989b] Kroch, A. (1989b). Reflexes of grammar in patterns of language change. *Language Variation and Change*, 1:199–244.

[Lightfoot, 1991] Lightfoot, D. (1991). *How to Set Parameters*. A Bradford Book. The MIT Press.

[Lupyan and Dale, 2009] Lupyan, G. and Dale, R. (2009). Language structure is partly determined by social structure. *submitted to Nature*. downloaded from http://psychology.stanford.edu/~jlm/pdfs/LupyanDaleSubNature.pdf on 6 Aug 2009.

[Mesoudi and Whiten, 2008] Mesoudi, A. and Whiten, A. (2008). The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Phil. Trans. R. Soc. B*, 363:3489–3501.

[Milgram, 1967] Milgram, S. (1967). The small world problem. *Psychology Today*, 2:60–67.

[Milroy and Milroy, 1985] Milroy, J. and Milroy, L. (1985). Linguistic change, social network and speaker innovation. *Journal of Linguistics*, 21(2):339–384.

[Mufwene, 2008] Mufwene, S. S. (2008). *Language Evolution Contact, Competition and Change*. Continuum International Publishing Group.

[Nettle, 1999a] Nettle, D. (1999a). Is the rate of linguistic change constant? *Lingua*, 108:119–136.

[Nettle, 1999b] Nettle, D. (1999b). Using social impact theory to simulate language change. *Lingua*, 108:95–117.

[Newman, 2002] Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters*, 89(20).

[Newman and Park, 2003] Newman, M. E. J. and Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E*.

[Newman et al., 2002] Newman, M. E. J., Watts, D. J., and Strogatz, S. H. (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99:2566–2572.

[Niyogi, 2002] Niyogi, P. (2002). *Linguistic Evolution through Language Acquisition*, chapter Theories of cultural evolution and their application to language change, pages 205–234. In [Briscoe, 2002].

[Niyogi, 2006] Niyogi, P. (2006). *The Computational Nature of Language Learning and Evolution*. The MIT Press.

[Niyogi and Berwick, 1995] Niyogi, P. and Berwick, R. C. (1995). The logical problem of language change. Technical report ai memo 1516/cbcl paper 115, MIT AI Laboratory and Center for Biological and Computational Learning, Department of Brain and Cognitive Sciences.

[Niyogi and Berwick, 1997] Niyogi, P. and Berwick, R. C. (1997). Evolutionary consequences of language learning. *Linguistics and Philosophy*, 20:698–719.

[Niyogi and Berwick, 2009] Niyogi, P. and Berwick, R. C. (2009). The proper treatment of language acquisition and change in a population setting. *Proceedings of the National Academy of Sciences*, 106(25):10124–10129.

[Palla et al., 2007] Palla, G., Barabási, A.-L., and Vicsek, T. (2007). Quantifying social group evolution. *Nature*, 446:664–667.

[Pearl and Weinberg, 2007] Pearl, L. and Weinberg, A. (2007). Input filtering in syntactic acquisition: Answers from language change modeling. *Language Learning and Development*, 3(1):43–72.

[Rafferty et al., 2009] Rafferty, A. N., Griffiths, T. L., and Klein, D. (2009). Convergence Bounds for Language Evolution by Iterated Learning. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society.*

[Reali and Griffiths, 2008] Reali, F. and Griffiths, T. L. (2008). Words as alleles: Connecting language evolution with Bayesian learners to models of genetic drift. *in preparation.*

[Reali and Griffiths, 2009] Reali, F. and Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, 111(3):317–328.

[Schneider, 2008] Schneider, E. W. (2008). Accommodation versus identity? A response to Trudgill. *Language in Society*, 37(2):262–267.

[Schwämmle, 2005] Schwämmle, V. (2005). Simulation for competition of languages with an ageing sexual population. *International Journal of Modern Physics C*, 16:1519.

[Smith, 2009] Smith, K. (2009). Iterated learning in populations of Bayesian agents. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society.*

[Toivonen et al., 2006] Toivonen, R., Onnela, J.-P., Saramäki, J., Hyvönen, J., and Kaski, K. (2006). A model for social networks. *Physica A*, 371(2).

[Travers and Milgram, 1969] Travers, J. and Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, 32(4):425–443.

[Troutman et al., 2008] Troutman, C., Clark, B., and Goldrick, M. (2008). Social networks and intraspeaker variation during periods of language change. In *University of Pennsylvania Working Papers in Linguistics.*

[Trudgill, 1982] Trudgill, P. (1982). Linguistic accommodation: Sociolinguistic observations on a socio-psychological theory. In Fretheim, T. and Hellan, L., editors, *Papers from the Sixth Scandinavion Conference of Linguistics*, pages 284–297. Tapir Trondheim.

[Trudgill, 2004] Trudgill, P. (2004). *New-Dialect Formation*. Edinburgh University Press.

[Trudgill, 2008a] Trudgill, P. (2008a). Colonial dialect contact in the history of European languages: On the irrelevance of identity to new-dialect formation. *Language in Society*, 37:241–280.

[Trudgill, 2008b] Trudgill, P. (2008b). On the role of children, and the mechanical view: A rejoinder. *Language in Society*, 37(2):277–280.

[Trudgill et al., 2000] Trudgill, P., Gordon, E., Lewis, G., and MacLagan, M. (2000). Determinism in new-dialect formation and the genesis of New Zealand English. *Journal of Linguistics*, 36(2):299–318.

[Walker et al., 2009] Walker, B., Fay, N., Rogers, S., and Swoboda, N. (2009). An experimental investigation of the role of collaboration in the evolution of communication systems. In Taatgen, N. and van Rijn, H., editors, *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

[Wang and Minett, 2005] Wang, W. S.-Y. and Minett, J. W. (2005). The invasion of language: emergence, change and death. *Trends in Ecology and Evolution*, 20(5):263–269.

[Watts, 2004] Watts, D. J. (2004). The "new" science of networks. *Annu. Rev. Sociol.*, 30:243–270.

[Watts, 2007] Watts, D. J. (2007). A twenty-first century science. *Nature*, 445:489.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393:440–442.

[Weinreich et al., 1968] Weinreich, U., Labov, W., and Herzog, M. I. (1968). *Directions for Historical Linguistics*, chapter Empirical Foundations for a Theory of Language Change, pages 95–188. University of Texas Press, Austin & London.

[Wray and Grace, 2007] Wray, A. and Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, 117:543–578.

[Yang, 2002] Yang, C. D. (2002). *Knowledge and learning in natural language*. Oxford Press.

[Yang, 2004] Yang, C. D. (2004). Universal grammar, statistics or both? *Trends in Cognitive Sciences*, 8(10):451–456.

A Matlab Code

A.1 Network Properties

```
function [c] = adj_cluster(a, d)
% clustering coefficient for directed(?) as well as undirected graphs
c = 0.0;
% only looping over vertices with degree > 1, most of the time it'll be all
    of them but we have to avoid divisions by zero..
for i=1:length(d)
  if d(i)>1
    % find adjacent vertices (incoming edges)
    x = find(a(:,i));
    e = sum(sum(a(x,x))); % count number of edges between pairs of adjacent
        vertices
    c = c + e/(d(i)*(d(i)-1));
  end;
end;
c = c/length(d);

function adj_to_dot(a, f)
f = fopen(f, 'w');
fwrite(f, 'graph {');
[j,k] = find(triu(a));
for i=1:length(k)
    fprintf(f, '%d -- %d;\n', j(i), k(i));
end;
fwrite(f, '}');
fclose(f);

function adj_to_dot_bw(a, h, f)
f = fopen(f, 'w');
write_dot_header(f);
[j,k] = find(triu(a));
for i=1:length(k)
    fprintf(f, '%d -- %d [penwidth=%d];\n', j(i), k(i), 1+(h(j(i)) == h(k(i))
        ));
end;
for i=find(h)
    fprintf(f, '%d [fillcolor=black,fontcolor=white]\n', i);
end;
fwrite(f, '}');
fclose(f);

function [d] = adj_to_deg(a)
% returns the (row) vector of IN-degrees of each node in the adjacency matrix
d = full(sum(a));

function write_dot_header(f)
fwrite(f, 'graph { node [shape=circle,style=filled,label=""]');
```

A.2 Small World Networks

```
function [a,d] = davidsen_create(n,m)
% n = total number of vertices
% m = maximum number of initial connections per vertex, m = 1/p from davidsen
    's model

% add first three vertices which are connected to each other
a = [not(speye(3)) sparse(3,n-3); sparse(n-3,n)];
d = repmat(2,1,n);

for i = 4:n
    % for every new vertex add 2 DISTINCT initial connections
    nb = ceil(rand(1,2).*[i-1 i-2]);
    nb(2) = mod(nb(1)+nb(2)-1,i-1)+1;
    % then make up to m-2 triadic closures (maximum degree is therefore m)
    [a,d] = davidsen_make_closures(a,d,i,nb,m);
end;

function [a,d] = davidsen_replace(a,d,i,m)
% replace vertex i of the given network

% remove vertex and update degrees of neighbours
d(find(a(i,:))) = d(find(a(i,:)))-1;
% erase links so that make_closures doesn't produce self-transitions
a(i,:) = 0;
a(:,i) = 0;

% add 2 DISTINCT initial connections
nb = [mod(i+cumsum(floor(rand(1,2)*(length(d)-1))),length(d))+1 zeros(1,m-2)
    ];
if i == nb(2)
    nb(2) = 1+mod(i,length(d));
end;
% then make up to m-2 triadic closures (maximum degree is therefore m)
[a,d] = davidsen_make_closures(a,d,i,nb,m);

function [a,d] = davidsen_make_closures(a,d,i,nb,m)
% select up to m-2 new neighbours for vertex i by triadic closures
% nb is a 1xm-vector with the first 2 values already set to the 2 initial
    neigbours
for j = 3:m
    % pick one of the current neighbours for the next closure
        nn = find(nb,ceil(rand*(j-1)));
    nn = nb(nn(end));
    % if the node has other neighbours make a closure
    if d(nn) > 0
        cl = find(full(a(nn,:)),ceil(rand*d(nn)));
        nb(j) = cl(end);
    else % otherwise reselect one of the previous neighbours
        nb(j) = nb(ceil(rand*(j-1)));
```

```
    end;
end;
a(i,nb) = 1;
a(nb,i) = 1;
d(i) = nnz(a(i,:));
d(nb) = d(nb) + 1;

function [a,d] = davidsen_evolute(a,d,m,g)
% run g complete generations
for i = 1:g
    for j = 1:length(d)
        [a,d] = davidsen_replace(a,d,j,m);
    end;
end;
```

A.3 Data Sampling & Inference

```
function [d] = sample_data_discrete(a, i, h, e, b)
% a = adjacency matrix
% i = agent receiving data
% h = hypothesis vector of all agents
% b = number of datapoints to sample
% e = probability of producing the 'wrong' signal
% returns the number of marked (1) datapoints
nb = find(a(i,:));
% sample b agents
nb = nb(ceil(rand(1,b)*length(nb)));
% sample 1 datapoint from each agent
d = rand(1,b) < e;
% flip datapoint if agent has h1
d(find(h(nb))) = 1 - d(find(h(nb)));
d = nnz(d);

function [h] = learn(pr, e, b, d, map)
% pr = prior
% e = error in signal production
% d out of b observations are from h1
% maximum a posteriori or sampling
p = pr.*[(1-e)^(b-d)*e^d e^(b-d)*(1-e)^d];
if map
    if p(1) == p(2)
        h = round(rand);
    else
        h = p(2) > p(1);
    end;
else
    % sampling from the posterior
    h = rand > p(1)/sum(p);
end;

function [h] = init_h(n, init)
```

```
% initialise hypothesis in right proportions and shuffle them
h = [zeros(1,n-ceil(init*n)) ones(1,ceil(init*n))];
h = h(randperm(n));

function [h,p] = init_h_p(n, init, g)
% initialise hypothesis in right proportions and shuffle them
h = init_h(n, init);
% vector that stores the distribution at each timestep
p = [nnz(h) zeros(1,n*g)];
```

A.4 Simulations

```
   function stationarydistributions(p,in)
n = [10 50 100 200]; % population sizes
m = [5 5 5 5]; % initial connectivity for small world networks
e = .05; % error term
b = 3; % bottleneck
g = 6000; % generations (rec. 6000 (includes snapshots at every 1000))
t = 50; % trials (rec. 50)

for i = 1:length(n)
    rand('twister', sum(100*clock));
    for map = 0:1
        dlmwrite(['stat/sw' num2str(n(i)) '-p' num2str(p) '-in' num2str(in) '
            -map' num2str(map)], stat_sw(n(i), m(i), in, [p 1-p], e, b, map, g
            , t));
        dlmwrite(['stat/full' num2str(n(i)) '-p' num2str(p) '-in' num2str(in)
             '-map' num2str(map)], stat_full_network(n(i), in, [p 1-p], e, b,
            map, g, t));
    end;
end;

function [s] = stat_full_network(n, init, pr, e, b, map, g, t)
% returns the stationary distribution of t trials after g generations each
shots = 6;
s = zeros(shots,t);
for k = 1:t
    h = init_h(n, init);
    s(1,k) = 0;
    for l = 1:shots
        for i = 1:g/shots
            for j = 1:n
                % sample agents
                a = ceil(rand(1,b)*n);
                % sample 1 datapoint from each agent
                d = rand(1,b) < e;
                % flip datapoint if agent has h1
                d(find(h(a))) = 1 - d(find(h(a)));
                h(j) = learn(pr, e, b, nnz(d), map);
            end;
            s(l,k) = s(l,k) + nnz(h);
```

```
        end;
        if l ~= shots % forward intermediate result
            s(l+1,k) = s(l,k);
        end;
    end;
end;
% normalise snapshots by generation proportion
s = s.*repmat(shots./[1:shots]', 1, t);
% normalize all by n
s = s./(n*g);

function [s] = stat_sw(n, m, init, pr, e, b, map, g, t)
% returns the stationary distribution of t trials after g generations each
shots = 6;
s = zeros(shots,t);
for k = 1:t
    h = init_h(n, init);
    s(1,k) = 0;
    [a,d] = davidsen_create(n,m);
    [a,d] = davidsen_evolute(a,d,m,1);
    for l = 1:shots
        for i = 1:g/shots
            for j = 1:n
                [a,d] = davidsen_replace(a,d,j,m);
                h(j) = learn(pr, e, b, sample_data_discrete(a,j,h,e,b), map);
            end;
            s(l,k) = s(l,k) + nnz(h);
        end;
        if l ~= shots % forward intermediate result
            s(l+1,k) = s(l,k);
        end;
    end;
end;
% normalise snapshots by generation proportion
s = s.*repmat(shots./[1:shots]', 1, t);
% normalize all by n
s = s/(n*g);
```

A.5 Visualisation

```
function [h] = hist3(x, z)
% 3d histogram
h = zeros(size(x,1), length(z));
for i = 1:size(x,1)
    h(i,:) = hist(x(i,:), z);
end;

function f = stationary_histogram(p, in, map)
ns = [2 10 50 100]; %2

d = 'data-new/';
```

```
f = figure();
for i = 0:1
    if i == 0
        pref = 'full';
    else
        pref = 'sw';
    end;
    for j = 1:length(ns)
        if or(j>1,i==0)
            data = dlmread([d pref num2str(ns(j)) '-p' num2str(p) '-in'
                num2str(in) '-map' num2str(map)]);
    %         d = [hist3(data, bins)./size(data,2) zeros(size(data,1),1);
        zeros(1,length(bins)+1)]
            plot_stationary(subplot(2,length(ns),i*length(ns)+j), data);
            title(['n = ' num2str(ns(j)) ', mean = ' num2str(mean(data(end,:)
                ))]);
        end;
    end;
end;

function [f] = figure_development(n, p, init, map)
m = 5;
p = [p 1-p];
e = .05;
b = 3;
g = 6000;

xg = [0:1/n:g];

f = figure();
for i = 1:2
    subplot(2,1,i);
    if i == 1
        [ps,cs] = dist_full_network(n, init, p, e, b, map, g);
    else
        [ps,cs] = dist_sw(n, m, init, p, e, b, map, g);
    end;
    ps = ps/n;
    plot(xg, ps, 'k-');
    hold on;
    ps = cumsum(ps)./[1:n*g+1];
    plot(xg, ps, 'k.');
%    communicative accuracy
%    plot(xg, cs, 'b--');
    if i == 1
        title(['development in a fully connected population (n = ' num2str(n)
            ')']);
    else
        title(['development in a small world network (n = ' num2str(n) ', m =
            5)']);
```

```
    end;
    axis([0 g 0 1]);
    xlabel(['generations (signifying complete population turnovers, i.e. 1
        generation = ' num2str(n) ' replaced agents)']);
    ylabel('proportion of population with hypothesis h_1');
    legend('current distribution in population', ['cumulative (stationary)
        distribution (' num2str(ps(end)) ')']);
end;

function f = properties_network

g = 500;
ns = [100 200];
m = 5;

f = figure();

subplot(1,2,1);
hold on;
xlabel('generations');
ylabel('clustering coefficient');

subplot(1,2,2);
hold on;
xlabel('generations');
ylabel('mean degree');

for n = ns

    c = zeros(1,g); % clustering coefficient
    k = zeros(1,g); % mean degree

    % add first three vertices which are connected to each other
    a = [not(speye(3)) sparse(3,n-3); sparse(n-3,n)];
    d = repmat(2,1,n);
    for i = 4:n
        % for every new vertex add 2 DISTINCT initial connections
        nb = ceil(rand(1,2).*[i-1 i-2]);
        nb(2) = mod(nb(1)+nb(2)-1,i-1)+1;
        % then make up to m-2 triadic closures (maximum degree is therefore m
            )
        [a,d] = davidsen_make_closures(a,d,i,nb,m);
        c(i) = adj_cluster(a(1:i,1:i),d(1:i));
        k(i) = mean(d(1:i));
    end;
    for i = n+1:g
        [a,d] = davidsen_replace(a,d,mod(i,n)+1,m);
        c(i) = adj_cluster(a,d);
        k(i) = mean(d);
    end;
```

```
    % clustering
    subplot(1,2,1);
    if n == 100
        style = 'k-';
        rstyle = 'k--';
    else
        style = 'k:';
        rstyle = 'k-.';
    end;
    plot(c, style);
    plot(k./[1:n repmat(n, 1, g-n)], rstyle);

    % mean degree
    subplot(1,2,2);
    plot(k, style);
end;
subplot(1,2,1);
legend('n = 100', 'random n = 100', 'n = 200', 'random n = 200');
subplot(1,2,2);
plot([1 g], [5 5], 'k--');
legend('n = 100', 'n = 200');
```